UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

Genetic characterization of the South Portugal population with GlobalFiler™ Express kit – internal validation and forensic applications

Carina Ema Silva Almeida

Dissertação

Mestrado em Biologia Humana e Ambiente

2014

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

Genetic characterization of the South Portugal population with GlobalFiler™ Express kit – internal validation and forensic applications

Carina Ema Silva Almeida

Dissertation advisors:

Professora Doutora Deodália Dias - Faculdade de Ciências da Universidade de Lisboa.

Especialista Superior em Medicina Legal Paulo Dario - Serviço de Genética e Biologia Forenses da Delegação do Sul do Instituto Nacional de Medicina Legal e Ciências Forenses.

Dissertação

Mestrado em Biologia Humana e Ambiente

2014

Nota prévia

Devido ao facto do Inglês ser a língua científica universal, a presente dissertação de mestrado encontra-se escrita na língua inglesa.

As referências bibliografias foram elaboradas segundo os parâmetros da revista científica *Forensic Science International: Genetics,* uma vez que esta é uma das mais relevantes na área onde esta tese se enquadra**.**

Acknowledgements

As previously stated, all of this thesis will be written in English, however I would like to acknowledge everyone that helps me to concluding this stage of my life in Portuguese.

Assim, aqui fica o meu *MUITO OBRIGADA* a todos aqueles que me ajudaram na concretização deste projecto.

Ao Professor Doutor Jorge Costa Santos e a Dra. Teresa Ribeiro por terem permitido que realizasse a minha tese no Serviço de Genética e Biologia Forense da Delegação do Sul do INMLCF.

À Professora Doutora Deodália Dias, minha orientadora interna, por toda a orientação, pela sua disponibilidade, por todo apoio e incentivo e principalmente por ter acreditado em mim e me ter concedido a oportunidade de trabalhar numa área tão incrivelmente fascinante como é a genética forense.

Ao meu orientador externo Mestre Paulo Dario, por todo o apoio e ajuda prestada ao longo deste ano de tese, especialmente na fase final e mais importante de todas que foi o tratamento dos dados estatístico e a escrita efectiva desta dissertação. Por ter confiado em mim e no meu trabalho e por, apesar do tão grande volume de trabalho, ter tido sempre algum tempo para me ajudar a resolver os problemas que sugiram ao longo dos tempos.

À Catarina Dourado por me ter "orientado" nos primeiros dias no laboratório, me ter ajudado a ganhar mão e ensinado tudo que seria necessário saber antes de começar efectivamente a trabalhar no laboratório. Por ter disponibilizado o seu tempo para participar também um pouco nesta tese fazendo o estudo de reprodutibilidade necessário e, acima de tudo, por me ter ajudado em tudo o precisei.

À Rita Oliveira Dario, por tal como a Catarina, me ter ajudado nos primeiros tempos de laboratório, por toda a disponibilidade, apoio e ajuda prestada, quer inicialmente na parte laboratorial quer na fase final de escrita desta tese.

A todas as “meninas da genética”, por terem ajudado estes dias/meses a passar mais rapidamente, por todas as gargalhadas, boa disposição e bons momentos partilhados.

E a todos os outros funcionários do Serviço de Genética e Biologia Forense da Delegação do Sul do Instituto Nacional de Medicina Legal, pela simpatia e disponibilidade.

A todos os meus amigos (vocês sabem quem são), e especialmente ao Tiago. Obrigada pela tua amizade, por todos os cafés tomados ao longo destes (muitos) anos, pelas longas horas de estudo de grupo (que na maioria das vezes eram de tudo menos de estudo), por todo o apoio e ânimo. Espero que a nossa amizade continue por muitos e muitos anos, até porque, inevitavelmente, estaremos os três ligados para sempre!

Ao André, por todo o apoio, encorajamento e por toda a paciência. Obrigada por me teres ajudado a passar por mais esta fase, por compreenderes a minha falta de tempo “para nós”, por alegrares os meus dias cinzentos e, acima de tudo, por estares sempre aí para mim quando preciso.

Aos meus irmãos, cunhada e sobrinhos pelo pouco tempo que tive para os programas de família neste último ano. Especialmente à minha irmã pelo apoio incansável, pela ajuda na organização, por me ter aturado em todos os dias maus e por ter tentado colmatar todas as pequenas falhas que cometi. Muito obrigada por tudo.

Por fim gostaria de agradecer aos meus pais por todo o esforço, paciência, apoio e encorajamento, sem o qual não teria conseguido chegar onde cheguei. Especialmente à minha mãe que apesar de já não estar cá para me ver concluir esta etapa, tenho a certeza que onde quer que esteja estará muito orgulhosa de mim. Li algures nessa imensidão que é a internet que “a maneira mais bela de recordarmos os outros é sermos aquilo que fizeram de nós”, obrigada por tudo mãe, especialmente por teres feito de mim a mulher que hoje sou.

Abstract

In forensic analysis, DNA is used in three major areas: to perform individual identification analyses, biological kinship testing and criminalistics. To perform these tests, STRs from DNA non-coding regions are currently used, amplified emplying commercial kits. In the last years, with the evolution of technology and the extension of the European Standard Set to 12 loci, new kits have been released to be used in this area. The GlobalFiler™ Express is one of them. It's a 6 dyes kit which permits direct amplification of 24 loci (21 autosomal STR loci, 1 InDel locus, 1 Y-STR locus and Amelogenin) in a single PCR reaction.

With the purpose of investigating the potential and limitations of using this set of molecular markers in South Portuguese population an internal validation study was performed, as well as a population study with 404 unrelated individuals involved in paternity casework, residing in the South Portugal area. The 404 bloodstains were directly amplified using GlobalFiler™ Express and the amplified products were separated and detected by capillary electrophoresis.

The internal validation study confirmed that the markers have all the requirements to be used hereafter in forensic casework. Allele frequencies of each marker were estimated using Arlequin V3.5 software and no deviation from Hardy-Weinberg equilibrium was found. Forensic parameters were analyzed using PowerStats V1.2. The SE33 was the most polymorphic locus and the TPOX was the least one. The combined power of discrimination was 0.99999999999999999999999981765, the combined probability of match was $1.8356x10^{-26}$ and the combined power of exclusion was 0.99999999966339800.

In conclusion, this commercial kit fulfills all the requirements for forensic identification use in South Portuguese population and its use can be an asset since it greatly reduces the time spent in analysis and greatly increases the power of discrimination and power of exclusion.

Keywords: Forensic genetics; Short tandem repeats (STR); GlobalFiler™ Express; DNA typing, Portugal

Sumário

Na genética forense a determinação do perfil genético de um individuo é utilizada principalmente em três grandes vertentes: identificação individual, investigação de parentesco biológico e criminalística biológica.

Para a determinação desse perfil genético os marcadores moleculares utilizados actualmente são do tipo microssatélites, os denominados STRs (do Inglês Short Tandem Repeats). Estes marcadores encontram-se dispersos por todo o genoma humano, contudo, os que são utilizados na área forense estão localizados apenas na zona não codificante do genoma. A escolha deste tipo de marcadores advém de inúmeros factores, nomeadamente pelo facto de possuírem um elevado poder de discriminação e de serem facilmente amplificados através da reacção de PCR.

Para além do uso de STRs autossómicos, em casos mais complexos ou em que o ADN se encontra degradado, também se recorre à determinação de perfil de ADN mitocondrial ou de marcadores do cromossoma Y ou X, para dar mais força à evidência biológica. Existem também estudos de perfis genéticos com marcadores do tipo InDel (polimorfismos de inserção/deleção) e de SNPs (do Inglês Single Nucleotide Polymorphisms). Contudo, estes últimos tipos de marcadores, individualmente, apresentam um menor poder de discriminação relativamente aos STRs e a nível de processos jurídico-criminais, em Portugal, não são ainda usados.

Tal como nos EUA existe o sistema CODIS, na Europa também é utilizado um painel de marcadores mínimo para identificação, o European Standard Set (ESS). Até 2009, esse painel era composto por 7 marcadores, tendo sido realizada nesse ano uma extensão do ESS, passando a partir daí a ser constituído por 12 marcadores. Esta extensão adveio da recomendação de incluir alguns marcadores do tipo mini-STR, pois estes possibilitam um aumento na probabilidade de obtenção de perfis em amostras degradadas, estabelecendo um maior número de marcadores comuns o qual possibilita a partilha de dados de perfis genéticos entre os diferentes países.

Após essa extensão, e com o enorme desenvolvimento de técnicas e tecnologias ocorrido nos últimos anos, actualmente têm surgido novos *kits* comerciais que permitem amplificar numa única reacção de PCR um elevado número de marcadores moleculares.

O GlobalFiler™ Express é um desses kits de nova geração, lançado em Setembro de 2012. Para além do facto de permitir que sejam amplificados 24 loci (D3S1358, vWA, D16S539, CSF1PO, TPOX, D8S1179, D21S11, D18S51, D2S441, D19S433, TH01, FGA, D22S1045, D5S818, D13S317, D7S820, SE33, D10S1248, D1S1656, D12S391, D2S1338, DYS391 e 1 Y InDel), numa única reacção de PCR, este *kit* permite a amplificação directa do ADN a partir da mancha de sangue ou da zaragatoa. Assim, e dado que a amplificação é extremamente rápida, ocorrendo em menos de 40 minutos, é possível a obtenção dum perfil genético, através de manchas de sangue, em pouco mais de hora e meia. Sendo também este o único *kit* no mercado que utiliza 6 fluoróforos na marcação dos diferentes loci, possibilita que exista espaço suficiente entre os marcadores adjacentes, minimizando os chamados OL (do Inglês *Off Ladder*) e, principalmente, permite que 10 destes marcadores sejam mini-STRs, com menos de 220 pares de bases.

Para avaliar o potencial e as possíveis limitações da utilização dos marcadores moleculares incluídos neste *kit* na população do Sul de Portugal, realizou-se a validação interna do mesmo e fez-se um estudo populacional utilizando 404 amostras de referência do serviço de genética e biologia forenses (SGBF-S), de indivíduos residentes no Sul de Portugal, envolvidos em casos de paternidade.

Através do processo de validação interno realizado, o *kit* GlobalFiler™ Express, demostrou preencher todos os requisitos necessários para amplificação e análise forense de ADN humano e, demonstrando características relevantes para que, futuramente, o mesmo possa vir a ser utilizado na rotina do SGBF-S do INMLCF. A sua especificidade, repetibilidade e reprodutibilidade foram comprovadas, assim como a sua concordância e capacidade para distinguir amostras contaminadas. Foi também realizado o estudo da sua sensibilidade, que não se revelou muito elevada, mas dado que

se trata de um *kit* para amplificar ADN a partir de amostras de referência, esse facto não deverá ser um entrave à sua utilização.

A nível populacional, usando o programa Arlequin V3.5, foram estimadas as frequências alélicas de cada marcador autossómico e verificou-se que todos os loci se encontravam em equilíbrio de Hardy-Weinberg. Seguidamente foi efectuado um estudo comparativo entre a amostra do Sul de Portugal e de outras populações. Os resultados dessa comparação permitiram aferir que as populações europeias em estudo (Portugal, Espanha, Itália, Alemanha, Áustria e Suécia) são muito similares entre si, assim como também a população caucasiana dos Estados Unidos América é muito próxima de todas elas. A única população que revelou diferenças significativas com a nossa amostragem foi a população da Coreia do Sul.

Para o cálculo dos parâmetros estatísticos de interesse forense foi utilizando o PowerStats V1.2. Individualmente o locus que apresentou maior polimorfismo foi o SE33 com 0,9457 de PIC, sendo também o locus com menor probabilidade de coincidência (0,0077), maior poder de discriminação (0,9923) e maior poder de exclusão (0,8788). Por outro lado o locus menos polimórfico foi o TPOX com um valor de PIC igual a 0,5806, sendo também este o que apresentou menor poder de exclusão (0,3300) e de discriminação (0,8182) e maior probabilidade de coincidência (0,1818). Com o uso combinado de todos os marcadores autossómicos, o poder de discriminação determinado foi de 0,99999999999999999999999981765, a probabilidade de coincidência foi de $1{,}8356 \times 10^{-26}$ e o poder de exclusão foi de 0,99999999966339800.

Pode assim concluir-se que este novo *kit* comercial preenche todos os requisitos necessários para uso em identificação genética na população do Sul de Portugal, sendo o seu uso uma mais-valia nesta área, uma vez que permite não só tornar todo o processo muito mais rápido como também aumentar, em muito, o poder de discriminação e de exclusão.

Palavras-chave: Genética Forense; Microssatélites (STR); GlobalFiler™ Express; Portugal

Index

List of Tables

List of Figures

List of Equations

List of Abbreviations

A	Adenine
AMOVA	Analysis of molecular variance
Bp	Base Pairs
C	Cytosine
CODIS	Combined DNA Index System
DNA	Deoxyribonucleic Acid
mtDNA	mitochondrial Deoxyribonucleic Acid
EDNAP	European DNA Profiling Group
ENFSI	European Network of Forensic Science Institutes
ESS	European Standard Set
FSS	Forensic Science Service
G	Guanine
F_{st}	Population pairwise genetic distance
He	Expected Heterozigoty
Ho	Observed Heterozigoty
HWE	Hardy–Weinberg Equilibrium
InDels	Insertion Deletion Polymorphisms
INMLCF-S	Instituto Nacional de Medicina Legal e Ciências Forenses, delegação do Sul
ISFG	International Society of Forensic Genetics
mtDNA	Mitochondrial DNA
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction

PD	Power of Discrimination
PE	Power of Exclusion
PI	Typical paternity Index
PIC	Polymorphism information content or power of information content
RFLP	Restriction Fragment Length Polymorphisms
RFU	Relative Fluorescence Units
SGBF-S	Serviço de Genética e Biologia Forenses – Delegação do Sul
SSR	Simple Sequence Repeat
STR	Short Tandem Repeat
SNP	Single Nucleotide Polymorphisms
SWGDAM	Scientific Working Group on DNA Analysis Methods
T	Thymine
VNTR	Variable Number of Tandem Repeat
W	Probability of Paternity

1. Introduction

1.1 The Deoxyribonucleic acid

The human genome is a complete set of our genetic information which is organized into deoxyribonucleic acid (DNA) molecules. The DNA is made of two long polypeptide chains compound by nucleotides. Each nucleotide is composed by a sugar molecule, a phosphate group and one more molecule called base. The four different types of nucleotides differ from each other only in the type of base, and so they can be abbreviated as A (adenine), T (thymine), G (guanine) and C (cytosine) according to the base contained. The DNA structure is a double stranded molecule, negatively charged, usually in a form of a double-helix (first described in 1953 by Watson and Crick). This double-helix structure is constructed based on nucleotides complementarity, A and T bases pair up as do the C and G bases, therefore, an A, for example, on one strand is paired with a T on the other strand. The DNA has two main aims: to contain the instructions needed to make proteins, so cells can build and maintain all the conditions needed for life and to create copies of itself to ensure that the same information is carryed when cells divide [1].

There are two types of DNA: nuclear, which is present in the nucleus cell organized in chromosomes, and mitochondrial DNA (mtDNA) a small portion of extranuclear DNA named as such due to being present in a cellular organelle known as mitochondria (Figure 1) [1,2].

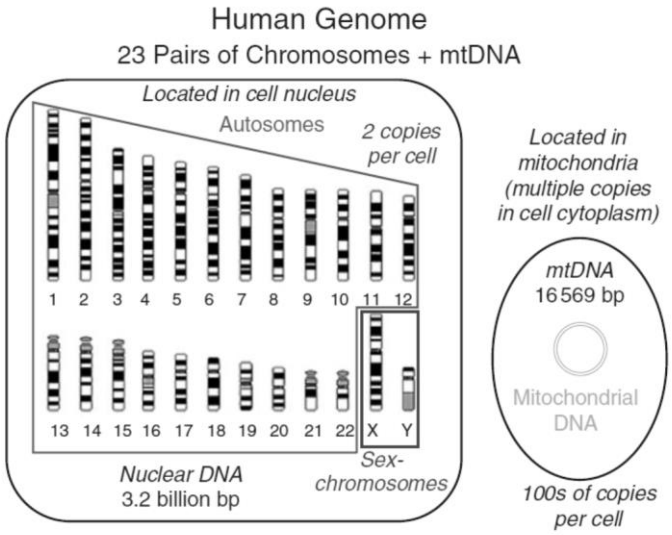


Figure 1 - The two types of DNA (nuclear and mitochondrial) present in every cell, which composes the human genome [3].

These two types of DNA have some differences. The nuclear DNA is packaged in 46 chromosomes: 22 pairs of homologous autosomal chromosomes, numbered from 1 to 22 according to their size, number 1 being the major chromosome and number 22 the smallest, and 2 sexual chromosomes, X and Y, that dictate the gender of the individual (a normal male has one X and one Y chromosome and a normal female has two X chromosomes). The latter, in a total, it has more than three billion base pairs (bp), while the mtDNA consists in a circular double stranded molecule with only approximately 16 569 bp in length [3]. The two types are passed through generations, although the nuclear DNA is inherited half from the mother and the other half from the father while the mtDNA is inherited exclusively from the mother, that is, there is no recombination and so all the members of the maternal lineage have the same information in their mtDNA [4–6].

The DNA material in chromosomes can be divided into coding and non-coding regions. The coding regions are called genes and contain all the information needed for encode and regulate proteins synthesis. Genes are only about 5% of the human genome being the remaining 95% non-coding regions. The function of this non-coding DNA regions is still unclear; it does not encode proteins, but contains numerous elements that are involved with genes' regulation, like promoters and repressors [2,3].

Is this molecule, present in all the cells of the body (except in red blood cells, which are enucleated), that contains all the information needed for life and, despite the basic information being equal in all the inviduals, it has some variations that are responsible for genetic variation and allow genetic identification. For this reason, and since its introduction in the middle 80's, the DNA typing has undergone numerous changes and has become an extremely important tool in the field of forensic sciences [7].

1.1.2 DNA in forensic science

According to the Locard Exchange Principle, each and every time a person makes contact with another person, object or place, a self-transfer occurs and pieces of evidences are left. This is one of the basic principles in forensic sciences: when a crime is committed some of that self-transfer can be used to recover DNA and link a person to a crime or a crime scene [8]. Regardless of these criminalistics analyses (biological traces), the DNA is currently used in forensic genetics to conduct individual identification (cases of missing persons and identification of corpses and skeletal remains) and in investigation of biological kinship (especially paternity tests) [4,5]. The type of DNA used when an analysis is performed depends on many factors (like sample type or state of degradation) but whenever possible nuclear DNA is preferred as it has more power of discrimination and allows for individual identification [4,5].

Obviously, when we talk about DNA in forensic science/genetics the first thing that is important to note is why this molecule is an exceptional source of information to identify an individual. DNA can be collected from any biological material, like blood, semen, hair, bones, skin and saliva, being the same regardless of the cell types that are used - all of cells have a common origin in a fertilized ovule and so the information is equal. There are some rare exceptions, such as individuals called chimaeras which possess two genetically distinct group of cells and produce different profiles from different tissues and as in individuals who recently received blood transfusion, bone marrow or organ transplant and that will produce a profile of the donor from those tissue; genetic

information normally does not change throughout a person's lifetime (however there are reports of allelic alteration in patients with some forms of cancer); DNA has high variations between individuals and, with the technologies used currently, only a small sample is needed to perform an analysis [6,8–10].

The first case resolved with the use of DNA fingerprinting was performed in England and it was an immigration case in the year of 1985. In the next year, in England too, DNA analysis was used as evidence for the first time in a criminal case: Colin Pitchfork was arrested after his blood sample genotype matched the one recovered from semen found at the crime scenes scenes [5,11].

Subsequently, DNA typing became a standard technique used in criminal investigations and made possible for thousands of cases to be solved and to the acquittal of many innocents wrongly convicted who might otherwise would still be in prison [5].

1.1.2.1 DNA mutations

The DNA molecule is very stable. However sometimes errors can occur during the replication and the DNA structure be changed. When these changes in the nucleotide sequence or arrangement of DNA are passed to the next generation, it is called a mutation.

Mutations can occur for a number of reasons, including chemical exposure, ultraviolet radiation, viral infection and background radiation. There are different types of mutations:

- Point mutations - a change from one DNA base to another. When the change is between two purines (A and G) or between two pyrimidines (C and T), it is known as transition. When the change is between a purine and a pyrimidine (A and C, A and T, G and C, or G and T), it is called a transversion.
- Insertion and deletions - mutations that produce insertions and deletions of DNA sequences;
- Chromosomal changes – type of mutations that affect large sections of the entire chromosomes. These can be inversions (a section of a

chromosome that ends up in reverse order) or translocations (when sections of a chromosome move to another chromosome) [12].

Mutations in the non-coding DNA regions lead to high levels of polymorphisms, because they are selectively neutral, i.e. do not confer benefit nor disadvantage in the capability of the individual to survive, and so, these DNA regions can accumulate mutations and the new alleles originated by those mutations may became abundant in the population originating new polymorphisms - natural occurring variants that are found at least in 1% of chromosomes in the general population [1].

1.1.2.2 DNA polymorphisms

The nuclear DNA sequence is approximately 99.9% indistinguishable among two individuals and only a small portion of DNA is answerable for the genetic variability between humans. And so, to perform an genetic identification, polymorphisms from these non-coding regions are analyzed [12].

There are two types of DNA polymorphisms: sequence polymorphisms and length polymorphisms. The sequence polymorphisms are usually a result of a point mutation that leads to change of a nucleotide. The length polymorphisms are the result of insertion or deletion of nucleotides, many times appearing in tandem repeats, differing the number of repeats between individuals on a locus (Figure 2) [3,10].

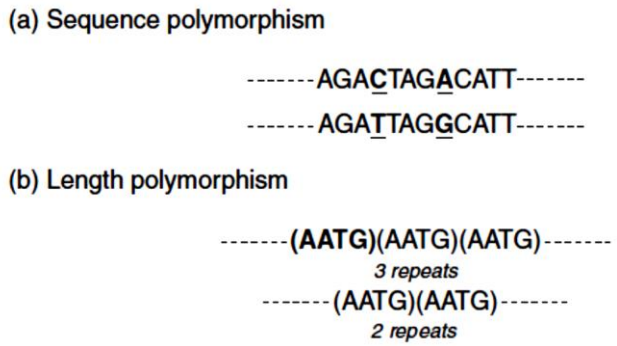


Figure 2 - Types of DNA polymorphisms: a) sequence polymorphism; b) length polymorphism [3].

1.2 Molecular markers for human identification

The term genetic/molecular marker is commonly used to refer to any gene or DNA sequence that has a known location on a specific chromosome. When a genetic marker is studied, a specific sequence is pursued and its location in the chromosome is designated as locus or loci (plural). The alternative forms of a gene or DNA sequence at a given locus is designated an allele [13].

As stated before, these specific sequences are transmitted by both parents, at each locus for each chromosome there is a specific sequence of paternal origin and specific sequence of maternal origin. If the allele located at each locus present the same structure, the individual is homozygous for that trait and when the alleles are different the individual is heterozygous. Thus the genetic profile of the individual is characterized by number of repeat units that have their genome in a given locus. Therefore, in the human genome, all of the chromosomes have a variety of unique repeat regions that can potentially serve for human identification [6,7,14].

1.2.1 Variable Number of Tandem Repeat

Firstly, it's important to refer that the many developments in molecular biology between 1960 and 1970 that enable the DNA sequencing analysis. In 1978, with the use of southern blotting, the DNA polymorphisms could be detected for the first time and, in 1980, the first polymorphic locus was communicated [2].

Throughout the history of DNA testing, the first molecular markers used for human identification were minisatellites also known as variable number of tandem repeats (VNTR) loci. In 1984, Alec Jeffreys was working in the laboratory of the Genetics Department in the University of Leicester when he discovered that some regions in DNA contained sequences of nucleotides repeated numerous times one after another and found that these patterns tend to appear in different arrays in different individuals [5,15].

These VNTRs loci are located in the chromosomes subtelomeric regions, having a core repeat sequence ranging from 8 to 100 base pairs long. To analyze the VNTRs a technique called restriction fragment length polymorphism (RFLP) is used, which was developed by Alec Jeffreys. The DNA is first digested by a restriction enzyme which cuts the regions of DNA surrounding the VNTRs followed by an agarose gel electrophoresis and then, a Southern blotting and probe hybridization which enables the detection of the polymorphisms (Figure 3) [2,3]. The technique that permits evaluation of the length variation of these repeats is known as genetic fingerprinting and was the first step to a technique presently named DNA typing [15,16].

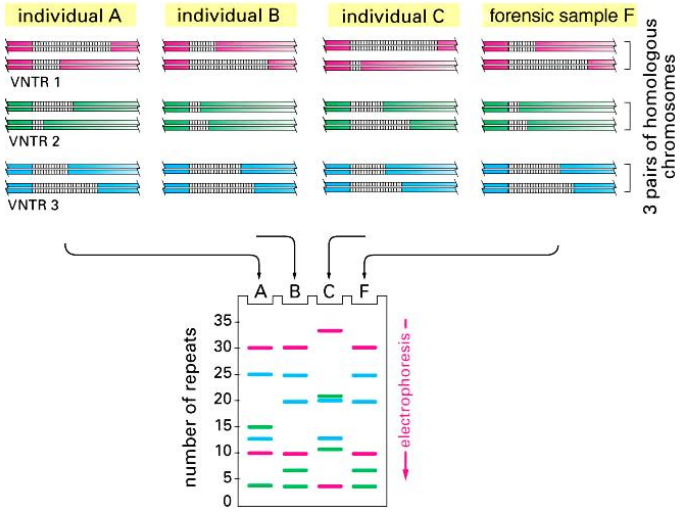


Figure 3 - Example of a VNTR analysis [17]**.**

This technique was used for many years and helped in solving many cases, despite its difficulties: perform a VNTR/RFLP analysis was not easy, it required a lot of time (many weeks), high quantities of DNA (limiting the number of samples that could be tested – it was not possible to use in degraded or trace samples). The results were difficult to interpret for the reason that they appeared as a bar code, and sometimes different labs would obtain different results with the same sample besides the fact that these markers have a very low discriminating power [1,18].

1.2.2 Short Tandem Repeat

In 1983, the polymerase chain reaction (PCR) – a process that permits amplifying only the region of interest by mimicking the DNA replication that occurs in cells – was conceptualised by Mullis and his collaborators. Some years after that, the analysis of short tandem repeats (STRs) was introduced in casework (middle of 90s) [1,19]. As it is represented in Figure 4, STRs polymorphisms aren't much different from VNTRs, the general structure is the same: it's still a length polymorphism, in which the variation between different allele is the number of repeated units. However, the STRs core repeat sequence ranging from 1 and 7 bp, and thus the amount of DNA needed to perform an amplification is smaller [2,3].

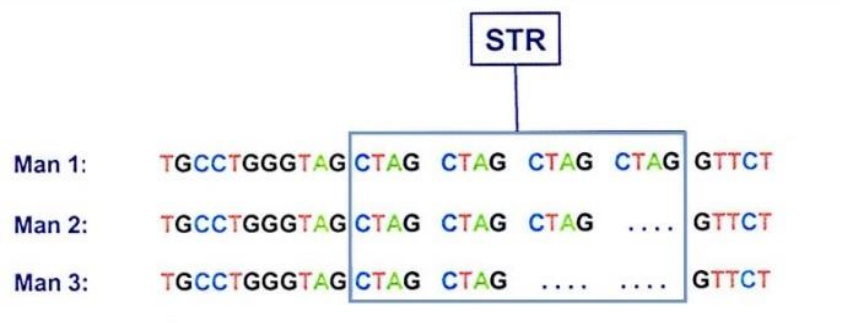


Figure 4 - Schematic representation of STRs polymorphisms (adapted [20]**).**

With the combination of PCR and STRs markers, the DNA analysis becomes a process that is less time consuming, with the possibility of analysing degraded DNA samples and the results are usually more easily interpretable [21,22]. Also, with the appearance of the automated fluorescent laser detection systems, it allowed for the analysis of STRs loci with similar size in one only assay. In the amplification, different fluorescent dyes are used for these loci and so, even with similar sizes, the software can distinguish the different loci. As result of that, the number of analysed allele have increased and so the power of discrimination has increased considerably too [18].

Nowadays, STRs are the typical markers used in forensic science, and for this reason, the STRs characteristics will be referred in more detail later.

1.2.3 Single Nucleotide Polymorphism

Another molecular marker that enables human identification in forensic genetic is single nucleotide polymorphisms (SNPs) (Figure 5). These makers occur in approximately every 1000 base pairs and most of them are present within a population as a result of a single mutation (single base variation which could consist in an insertion, deletion or a replacement of a single nucleotide) that occurred in a certain chromosome and thereafter spread across the population [2,23].

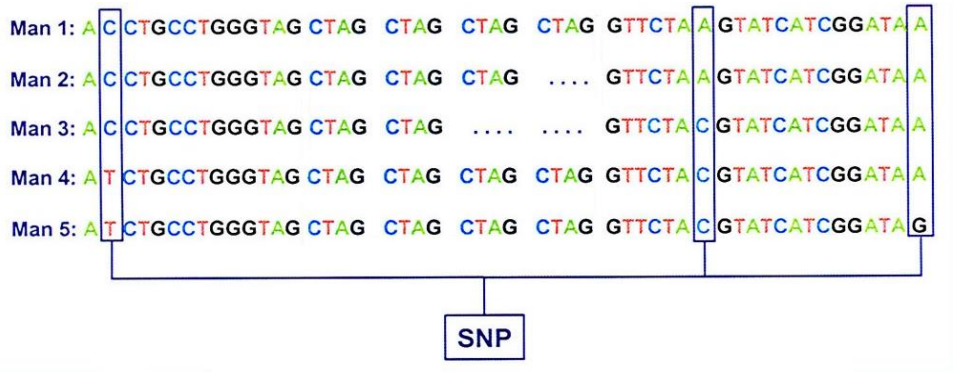


Figure 5 - Schematic representation of a Single Nucleotide Polymorphism (adapted [20]**).**

SNPs are very helpful in degraded samples because they can achieve amplicons about 40-50 bp, although, bi-allelic SNPs (SNPs normally just have two allele) are less polymorphic than STRs and so, the discriminated power is not as powerful as a full STR match (50-80 SNPs would be required to reach the power of discrimination achieved by 10 STRs) [24,25]. Besides that, SNP assays involves complex genotyping protocols, which are more time consuming, because they includes more steps, and are more expensive with the implementation of new techniques and high-throughput technologies [26].

1.2.4 Short Insertion Deletion Polymorphisms

Recently, a new approach has been introduced with a big potential to be used in challenging DNA cases: short insertion deletion polymorphisms or InDels markers. It's a type of genetic variation, very frequent and widely spread across the genome (hundreds of thousands), which are length polymorphisms created by insertions or deletions of one to thousands of nucleotides (figure 5).

It's originated from one single mutation event, occurring at a low frequency and is subsequently stable. InDels have a big variety in length difference among alleles, (reaching hundreds of kilobase pairs in extreme cases). Approximately half of InDels are bi-allelic, which is the presence or absence of the insertion or deleted segment, the other half are multiallelic owing to variable numbers of a segment of DNA that is repeated in tandem at a particular location [27,28].

Figure 6 - Schematic representation of an insertion deletion polymorphism (adapted [29]**).**

In forensic studies, InDels presents a large number of advantages such as low mutation rates, easy interpretation, small amplicons (less than 160bp), easily typed and the possibility of using multiplex PCR with a complete profile obtained with approximately 300 pg of DNA [26,30]. These facts make InDels ideal for the typing of degraded DNA samples and they can be used too in a lot of studies with a wide range of analysis and different purposes, like individual identification or ancestry, for example [12].

1.2.5 Chromosomes X and Y analysis

As an alternative to autosomal DNA, the study of Y and X chromosomes can be performed (with STRs, SNPs and InDels markers), especially in complex paternity testes, when a DNA sample from one of the parents is not available, in sexual assault, when a mixture of female and male sample appears and as complementary analysis of other genetic markers when a complete profile can't be obtain [30]. However, the discrimination of Y and X chromosome analysis is very low and as a result of that, in forensic science, Y and X-STRs are only used in special circumstances having only a minimal role in this area [9,31,32].

1.2.6 Mitochondrial DNA

Finally, and as noted before, there is another type of DNA, present in mitochondria, which can also be used in forensic genetics. The mtDNA is very useful in cases where it's not possible to obtain a profile with nuclear DNA, the reason being, there are thousands of copies of mtDNA per cell and so the probability of obtaining a DNA profile using mtDNA is higher due to be more likely that some of the copies will survive [2,30].

Besides the fact that mitochondrial DNA, like the Y markers, is a lineage marker, in other words, the information is passed through generations without suffering any type of recombination, which means that the power of discrimination is very low compared to STRs markers, the analysis of mtDNA is much more complex and time consuming. To analyze mtDNA is necessary to study the three hypervariable regions (HVS-I, HVS-II and HVS-III), because those are the most polymorphic regions and contain the highest levels of variation in the mtDNA. The end result of the sequencing is compared to the Cambridge Reference Sequence and the differences found between those two sequences permit to determine the samples haplogroup. The haplogroup enables to infer about geographical ascendance of a person or of a samples donor being tested and, in comparison with another person presumably of the same maternal linage, identify if the person in test belongs to the same linage [2,33].

1.3 Short Tandem Repeats: the excellence marker in forensic genetics

As stated previously, Short tandem repeats, *microsatellites* or *simple sequence repeats* (SSRs), are highly polymorphic regions of tandemly repeated DNA segments found surrounding the chromosomal centromere, throughout the human genome (approximately 500.000 STRs) that vary in length (through insertion, deletion or mutation), with a core repeated DNA sequence between 1 and 7 base pairs (bp) and the repeats typically range from 50 to 300 bp [1]. STRs markers occur on average once every 10 000 nucleotides and, with only approximately 3% of the total human genome accounting for microsatellites , thousands of polymorphic STRs have already been characterized in human DNA [3].

1.3.1 Types of STRs markers

The repeat motifs or patterns in STRs also differ by the length of the repeat unit: dinucleotide repeats are composed of two nucleotides tandemly repeated, trinucleotides by three, tetranucleotides by four and so on. In the human genome, dinucleotide loci are the most commonly, however, the tetranucleotide repeats are the most used in forensic genetics because they have less amplification artefacts, like stutters (amplicons that are one or more repeat units less in size then the true allele) and occur throughout the genome in all the chromosomes [4,30,34].

Besides the length of the repeat unit, STRs can be classified into different categories according to their structure:

- Simple Repeats - Repeat units all identical in length and sequence throughout;
- Compound Repeats - Two or more distinct adjacent simple repeats;
- Complex Repeats - Repeats of variable length or sequences.

In addition to these three types there is another STR structure that is named complex hypervariable repeat: that is a common tetranucleotide repeat

motif and some variant mono, di-, tri- and tereanucleotides are also scattered through the locus, and for this reason, loci with this structure are very polymorphic [35].

There are also STR alleles (even with simple repeats) which may contain some form of sequence variation compared to the more commonly observed. They have an incomplete repeat unit and are often called microvariants (e.g. allele 9.3 of TH01 locus contains nine tetranucleotide repeats and one incomplete repeat of three nucleotides) [4].

1.3.2 Application of STRs in Human identification: required characteristics

As previously stated, STRs were introduced to forensic casework in the mid-1990s. The first cases that used STRs typing to performed Human identification were in skeletal remains, the first one was from a murder victim and the next one allowed identification of Dr. Josef Mengele, the Auschwitz “Angel of Death”, with a femur from the exhumed body (found in Brazil in 1985) that was compared to the DNA from of Mengele's son and wife [36,37].

The STRs markers, when applicable to Human identification must have several characteristics like:

- high discriminating power;
- robustness and reproducibility of results when multiplexed with other markers;
- must conduct to a successful analysis of a large range of biological material;
- have low mutation rate;
- be suited for analysis of degraded DNA samples;
- make results easier to interpret and compare in computerized DNA databases [3].

Moreover, the selection of STR loci used in routine typing nowadays are typically chosen from separate chromosomes or if in the same chromosome far spaced to prevent linkage problems, are located essentially in non-coding

regions, readily amplified by the polymerase chain reaction, have minimal associations with diseases and high heterozygosity [38].

In forensic genetics, the combination of STRs used in commercial multiplexes nowadays are a combination of traditionally selected ones and the ones recommended by national and international organisations based on experimental data from many experimental studies on small multiplex typing kits designed by The Forensic Science Service (FSS) or by Caskey markers, in commercially available products from the major reagent suppliers and in recommendations by the Interpol DNA working in 1998 [9,39]. Currently, there is a larger number of commercial kits that permit easily amplification of a large number of STRs loci in a single run, saving time in analysis and providing a more sensitive and highly discriminating power, not only in unrelated persons but also within families [7,22].

1.3.3 Analysis of degraded DNA: reduced-sized STRs (Mini-STRs)

In many situations, when DNA is exposed to environmental conditions, like water, oxygen, ultraviolet irradiation or heat, it tends to break down and degrade. Those degraded DNA samples appear commonly in forensic casework and these samples are particularly difficult to analyse because only a small sample of intact DNA is present. In these situations, a loss of signal appears with the use of STR loci where the major fragments aren't amplified. In order to make degraded DNA investigation possible, analysis of mitochondrial DNA hypervariable regions is typically conducted, as the size of intact DNA molecules needed to perform amplification is shorter than the size needed for the others DNA molecular markers (Figure 7) and because the circular nature of mtDNA makes it more resistant to degradation than nuclear DNA [21,40–43]. However, as previously noted , mitochondrial DNA analysis is not as powerful for identification purposes as a full STR match [40].

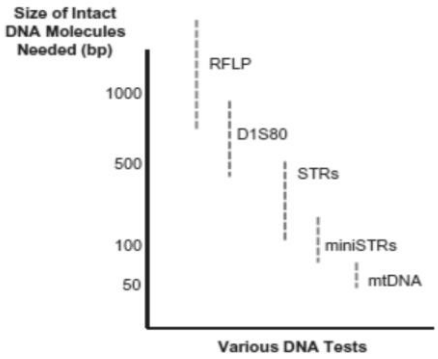


Figure 7 - Required DNA fragment sizes for the various DNA Tests [30]**.**

As a response for this problem a solution has been found: reduced-sized STRs, i.e., mini-STRs - which the amplification product have less than 150 bp [4,42]. In 1995, a first report of the success rate with degraded DNA samples using mini-STRs came from the analysis of victims of the Waco Branch Davidian fire and in the next years the success of these mini-STRs have been confirmed by innumerous publications like, for example, Wiegand and Kleiber in 2001 [44]. These authors confirmed that highly degraded samples or very low quantities of DNA could be more successfully typed when new redesigned PCR primers, as close as possible to the STR repeat region, were used. So, when the forward and reverse PCR primers are moved closer to the STR repeat regions (Figure 8), the product sizes are reduced while the same information is retained and the database compatibility could be maintained with convicted offender samples processed using commercial STR multiplexes [14,41,45,46].

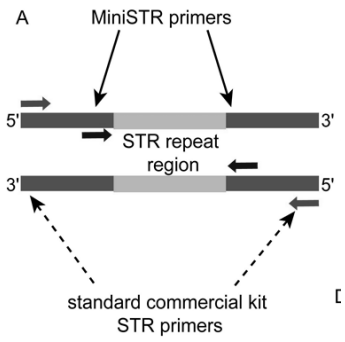


Figure 8 - Comparison of STR and Mini-STR primers insertions [46]**.**

Intra and inter-laboratory assays proved that reduced-size STRs have more success rates in recovering information from compromised DNA samples than conventional STR kits. With the use of mini-STRs, telogen hair shafts which containing very little nuclear DNA were successfully typed and enabled the identification of some of the World Trade Center victim using DNA extracted from burned and extremely damaged bones [30,47].

Therefore, since 2005, the European community recommends the inclusion of mini-STR loci in STR commercial kits because, when mini-STRs are used, the sensitivity of DNA detection increases dramatically and the opportunity to obtain a full DNA profile from compromised samples increased too [14,30,45].

As a result of the inclusion of mini-STRs in commercial kits, more degraded samples can be typed and many previously problematic materials and seemingly unsolvable human identity cases by using conventional multiplex amplification may now be resolved [14,24].

1.4 European Standard Set

The standard set used in Europe appears at the year of 1999, when the DNA working group of the European Network of Forensic Science Institutes (ENFSI) decided preceding a European Standard Set (ESS). This set includes seven loci: TH01, vWA, FGA, D21S11, D3S1358, D8S1179 and D18S51, which have been confirmed, in 2001 by a resolution of the European Council [39].

In 2005, the Prüm Treaty was established between Belgium, Germany, Spain, France, Luxembourg, Netherlands and Austria in order to intensify and accelerate the exchange of information between authorities and to allow comparisons between a certain DNA profile and the profiles recorded in existing automated data in the Member States [40]. In the same year, a recommendation was made by the European Forensic community to develop new STR multiplexes to provide greater discrimination power and enhance performance including newly introduced STRs that would extend the ESS loci,. The intention of this extension was to improve discrimination power, improve sensitivity of testing in reduced samples of DNA and increase the robustness and quality of the results [39].

Approximately at the same time, in 2006, the European Network of Forensic Science Institutes DNA working group and the European DNA Profiling group recommended evolution of DNA databases through Europe. For this evolution, there was the need for data sharing between countries and in order to facilitate the process more loci needed to be added in the standard set [40].

After all these recommendations and with the confirmation that mini-STRs are a very helpful tool in the analysis of degraded DNA samples, the extended European Standard Set was formally adopted in 2009, when the European Network of Forensic Science Institute voted in five additional (mini) STR loci (D12S391, D1S1656, D2S441, D10S1248, and D22S1045) to join to the already current seven [38].

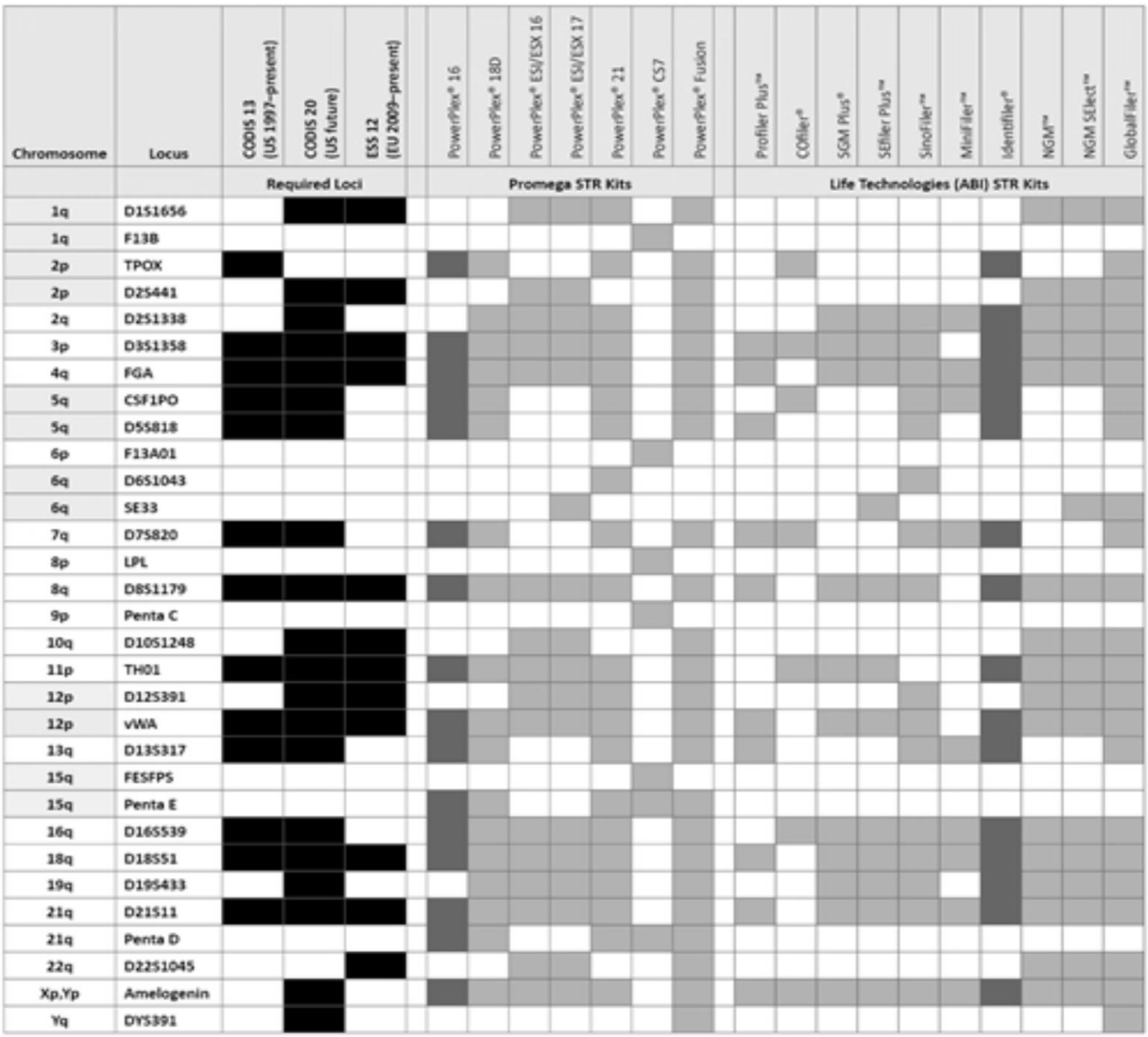
1.5 The evolution of STRs multiplexes

In the practice of forensic science, the laboratories rarely use in-house STRs assays, they buy quality-controlled commercial kits with allelic ladders, positive control samples, and premixed reagents including the fluorescently primers that target the specific locations in the human genome to be amplified by the polymerase chain reaction [38].

DNA testing technologies have had a lot of developments in the last two decades evolving quickly for more sensitive, rapid, and accurate measurements of STRs allele. At the same time, the number of STRs that can be simultaneously amplified has increased considerably [3]. The first commercial kit was developed in 1994, by Promega Corporation (Triplex CTT STR Multiplex System) and was able to amplify 3 STR markers: CSF1PO, TPOX, and TH01 [48]. After that, the Forensic Science Service (FSS) developed a STR typing system designed for forensic analysis: the “quadruplex” that could amplify four STR loci (TH01, FES/FPS, vWA and F13A1) with a matching probability of 1 in 10.000. In 1996, a six-locus STR system, with the amelogenin (the sex-determining marker) was introduced - second generation multiplex - and, with the used of complex STR D21S11 and FGA, the match probability decreased to 1 in 50 million. Subsequently, in 1999, a new multiplex with 10 STRs (Amp*Fl*STR® SGM Plus™) became available, which changed the probability of a match, between two unrelated individuals, to approximately 10^{-13} [1,2,30,49].

Also, in the last few years, different approaches to sequencing chemistry and developments in technology have led to next-generation sequencing (NGS) technologies and some commercial companies have developed a series of multiplexes, called next-generation kits, that are used actually by most laboratories around the world (Table 1). These next-generation kits, as PowerPlex® 16 HS, Amp*Fl*STR® Identifiler Direct or Plus and Amp*Fl*STR® NGM™ and now the GlobalFiler™ Express, contain enriched buffers and demonstrated the capability to generate STRs profiles even in the presence of polymerase chain reaction inhibitors. This enrichment has allowed for some of those new kits to perform direct PCR from bloodstains or buccal swabs, without extraction and purification of DNA, saving a lot of time in analyses [2,30,49].

Table 1 - Some of Commercial STR multiplexes kits used in forensic laboratories and the loci amplified by each one [7].



After the extension in the ESS and as a response to the ENFSI/ European DNA Profiling Group (EDNAP) recommendation to increase the number of STR loci available, new profiling kits become available with different combinations of 24 autosomal STR loci: CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D2S1338, D19S433, Penta D, Penta E, SE33, D1S1656, D12S391, D2S441, D10S1248, D22S1045 and D6S1043. Those commercial kits are used worldwide, allowing the opportunity for share data across a large range of jurisdictions and allows the development of national DNA databases [7,21,40].

Some new kits have been recently released, like the GlobalFiler™ Express, including some new STR loci (D12S391, D1S1656, D10S1248,

D2S441, D22S1045 and SE33) that still are relatively unfamiliar for most forensic laboratories and need to be analysed in order to evaluate the behaviour of these markers in a particular population and their usefulness for forensic purposes.

1.5.1 The GlobalFiler™ Express kit: innovations and advances

The GlobalFiler™ Express was a new STR multiplex released in September 2012. It's a 6 dyes commercial kit that amplifies in a single PCR reaction a total of 24 loci. The dyes used to label samples are 6-FAM™ (emits blue), VIC® (green), NED™ (yellow), TAZ™ (red) and SID™ (purple). The sixth dye is LIZ® (orange) and is used to label the internal size standard - GeneScan™ 600 LIZ® Size Standard v2.0. The schematic representation of these 24 markers and the respective dyes are present in Figure 9 [50].

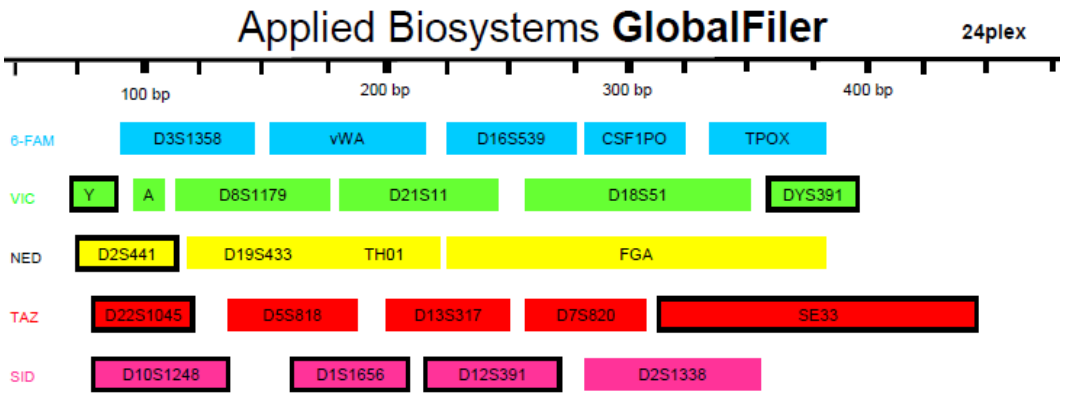


Figure 9 - Schematic representation of all molecular markers presents in the GlobalFiler™ express kit with the correspondent dye and amplicon size. The highlighted markers are the ones that are not present in AmpFlSTR® Identifiler kits. [51].

The 6 dyes configuration is a very important innovation as holding all the 24 loci in a 5 dyes system would result in numerous tradeoffs, like insufficient spacing between adjacent markers and the inclusion of ten mini-STRs would not be possible. Thus a 6 dyes allows an optimal performance, data recovery and genotyping accuracy [51].

The 24 loci included in this kit are 21 autosomal STR loci (D3S1358, vWA, D16S539, CSF1PO, TPOX, D8S1179, D21S11, D18S51, D2S441, D19S433, TH01, FGA, D22S1045, D5S818, D13S317, D7S820, SE33,

D10S1248, D1S1656, D12S391, D2S1338), 1 Y STR locus (DYS391), 1 Y insertion deletion locus (from the long arm of the Y chromosome) and amelogenin, the sex-determining marker [50]. All of the loci are tetranucleotide short tandem repeats except the D22S1045 which is a trinucleotide; ten are mini-STRs and five of them (D12S391; D1S1656; D2S441; D10S1248; D22S1045) are those which were adapted in 2009 to expand the European Standard Set to 12 and were recommended too by the ENFSI and EDNAP to be adopted for the analysis of degraded DNA samples, to improve the power of national databases and standardize the procedures across Europe [50,52,53].

A total of 589 distinct allele can be assign (343 total allele more 246 virtual bins that increase genotyping efficiency). The great number of STRs included in this kit permits a major compatibility across databases, the reason being that it's the only kit that contains all markers used in Europe and the markers recommended for inclusion in Combined DNA Index System (CODIS), reducing the risk of incorrect matches and increasing the power of discrimination about 9 orders of magnitude: preliminary calculations based on US Caucasian Databases determine that the Probability of identity (PI) of the GlobalFiler^{TM} Express is 7.12 x 10^{-26} (1 in 1.4 x 10^{25}) that is 10^{14} times greater than the total population that has ever lived on Earth (1.1 x 10^{11}) [50].

Besides the great power of discrimination, this kit save a lot of time in analysis because allows for direct amplification in less than 40 minutes for single source samples. More, the fact that the largest product of this kit amplification is inferior to 460 bp and the 10 mini-STRs have amplicons smalls than 220 bp, increases the chance of typing degraded samples or trace DNA [50].

The inclusion of the Y STR locus DYS391 is an important factor too since this marker has been proposed to help sex-typing confirmation when a deletion of amelogenin Y appears (in those cases only the amelogenin X allele is amplified and consequently a male sample appears as a female) [38]. This Y STR is a very stable locus and it's located more than 7 mega base pairs from amelogenin, which avoid deletions affecting AMEL Y and so it can type the gender correctly (Figure 10) [30]. In degraded samples, where the presence of DNA is so diminutive that none of those two markers (amelogenin and DYS391)

can be typed, the GlobalFiler^{TM} Express kit also has another marker in the Y chromosome: an Y InDel that can be helpful as this is a very little molecular marker (less than 90 bp) that needs a very few intact DNA molecules to be amplified [38,50,54].

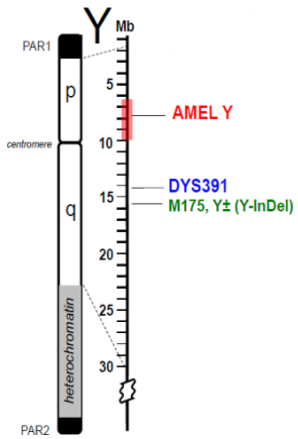


Figure 10 - Location of AMEL Y and the region around that can sometimes be deleted (red). The location of DYS391 (blue) and Y-InDel (green) are far enough to avoid deletions that affect AMEL Y (adapted [54]

In forensic analysis any STR used should give an identical profile regardless of the individual or laboratory that realized the analysis. It's this standardization that made possible comparing results between laboratories. Therefore, like any new technology that is applied in forensic casework, these 24 loci need to be rigorously evaluated throughout a validated program before they start being used with the South Portuguese population [30].

1.6 Validation Studies

The International Organization for Standardization (ISO) emits guidance documents on a large number of issues. The General Requirements for the Competence of Testing and Calibration Laboratories is present on ISO 17025:2005 and many DNA testing laboratories search for accreditation through this standard. Laboratory accreditation involves numerous steps and results from a successful inspection, by an accrediting body, which evaluates not only the equipment and technical procedures but also the training of the technical staff, the caseworks reports and the supporting documentation. Accreditation requests that the laboratory demonstrates and preserve good lab practices and to confirm the analyst work quality, proficiency tests are request, while methods and instruments are verified through validation processes [33].

Validation of the new methods is not only a step to achieve accreditation, it's a vital process in maintaining high-quality results, even in non-accredited labs. So, when new techniques or technologies are introduced in a laboratory, they must complete a validation process. This validation confirms that the method is suitable for a specific intended used and provides confidence in the obtained results [30,33].

When a validation process reaches its terminus, the method must be considered robust, that is, guarantees successful results in a high percent of the time; reliable (obtain results that are correct and reflects the samples being tested) and reproducible (the same results are obtained, each time a sample is tested) to be used hereafter. And as any result obtained in forensic laboratories needs to be valid in courts of law, it is the validation studies that provide the information about the limitations and specificities of a particular method, that will offer the necessary confidence as well as support the quality assurance in a particular lab [30,55].

There are two types of validation processes: developmental validation and internal validation. The first one involves testing the new STR loci or kits, new primer sets and new technologies for detecting STR alleles and is performed by the commercial company, the laboratory or laboratories that developed the new technique or technology. This developmental validation will

be provided in the manufacturer's manual. Internal validation, on the other hand, is performed by a laboratory which acquires the new method with the intention of verifying if the established procedures (by the development validation) will work efficiently in their laboratory [30].

In order to promote proper validation studies across the Europe, standardization procedures need be defined, however that standardization is not yet reached and nowadays the most accepted procedure, to make those validation studies, is the Scientific Working Group on DNA Analysis Methods (SWGDAM) recommendations [30,55,56].

1.6.1 Internal Validation Studies

As mentioned before, the internal validation is part of the implementation of a new method or technique in a lab, in order to understand if established procedures will work effectively in a specific laboratory. Internal validation studies typically include tests to measure sensitivity, precision, reproducibility, mixture analysis and nonprobative casework samples [33]. The SWGDAM recommends that each laboratory must determine the appropriateness of each study based on the technique and may determine whether performing certain tests is necessary or not. They should also develop standard operating procedures, providing guidelines that enable a successful completion of the experiments [30,56].

After the internal validation tests are complete, the new method can be introduced in the routine casework and the obtained results (technical procedures and guidelines for data interpretation) must be carefully documented, being used in case doubts appears in the routine casework [30].

1.6.1.1 PCR Amplification optimization

Before any use of the new methodology it is necessary to confirm if the reactions conditions offer the necessary degree of specificity and robustness. The amplification cycle number is focused on the performance of the new multiplex varying the optimal conditions supplied in the protocol to verify what is the optimal cycle number (producing sufficient amplified product) according to the technology used in the laboratory [3,56].

1.6.1.2 Species specific Study

In forensic casework, the DNA may appear in different context and might be contaminate with different biological sources. This test verifies if non-human DNA interferes with the ability to obtain reliable results; It allows to prove that the new kit will be able to identify and unequivocally characterized Human DNA without any interference from other animal species [3,56].

1.6.1.3 Reproducibility Study

The DNA tests results must be comparable between different laboratories, across distance and time. This test will express the precision (degree of mutual agreement between innumerous individual measurements, values and/or results) of the method performed in different conditions, to confirm that a same sample always has the same STRs profile. The reproducibility test is performed by another collaborator in another day, using the same protocol methodology [3,56].

1.6.1.4 Repeatability Study

In this study, the precision and the accuracy of the method performed in identical conditions was tested. It presents a good repeatability when the difference between two different results, for the same individual, is the same [3,56].

1.6.1.5 Contamination Study

In forensic casework, evidence samples are often from more than one contributor, becoming essentials that the typing system has the capability to detect mixtures. So, with this test, the ability of the DNA typing system to identify various components of mixture samples is study. A good method to perform human identification has the ability to discriminate mixtures samples from single source samples. Beside this test, negative controls must also be analyzed every time amplification is performed with the real samples, to avoid possible contaminations [3,56].

1.6.1.6 Sensitivity study

Sensitivity of a method is allows us to know the minimum quantity of genomic DNA needed to perform an analysis with reliable results. This test should determine minimal amount of DNA needed to obtain a reliable result and concentration range that enables the best results with this kit [3,56].

1.6.1.7 Concordance Study

This study is performed to verify if a given sample, with a known profile previously determined by a different PCR amplification system, will have the same genotype with the new PCR commercial multiplex kit [3,56].

1.6.2 Population Studies

Nowadays, genetic studies evolve the study of populations. In general, populations are a group of individuals that share a common ancestry. However, when we talk about population in genetics it's refers to a group of individuals residing in a specific area at a specific period of time. Thus, populations studies, in this context, are the study of inheritance of variations in specifics trails, in time and space with the intention to quantify the variation observed among different populations groups or inside a specific population [33].

The genetic diversity of a population is reflected in the number of allele of a particular genetic marker. The variability is greater when the number of

alleles of a marker is greater too. Among the factors that affect the genetic diversity of populations are migration, mutations and natural selection, being that closed and/or isolated populations that exhibit low levels of heterozygosity [3].

Population studies are part of the procedures needed to performed internal validation, as this data will enable the study of allele frequencies (number of copies of an allele in a population divided by the total number of all allele in the population in study). That data will be used in reporting population statistics and calculating the most part of the forensic parameters [3].

1.6.2.1 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium (HWE) predicts how gene frequencies will be inherited from generation to generation. Assuming that there is no genetic drift, mutation, gene migration or selection, in a large population with random breeding, the allelic frequencies will remain the same from generation to generation. Thus, when we have a molecular marker with two allele, A and a, the frequency of expected genotypes AA, Aa and aa, are respectively, p^2, 2pq and q^2. The equation for calculate HWE and the frequencies of AA, Aa and aa can be illustrated by the Punnet square (Figure 11), where p and q represent the frequencies of allele A and a [3,13,57].

When the observed genotypes are in agreement with the expected frequencies it is said that the population is in Hardy-Weinberg equilibrium and so allele and genotype frequencies will remain unchanged through time [13].

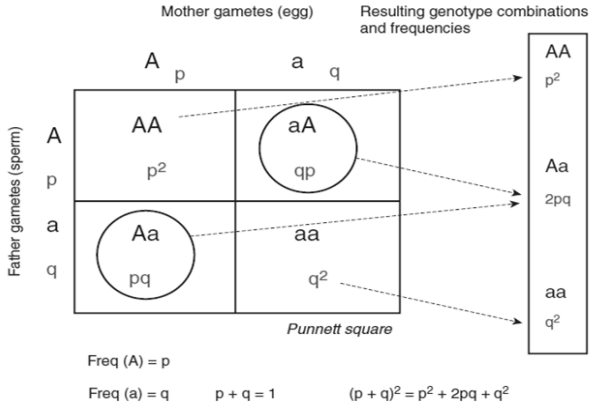


Figure 11 - Punnett square [3].

1.6.2.2 Population pairwise genetic distances

Population pairwise genetic distance (F_{ST}) is a measure of the difference in the allele frequency between two populations. The F_{ST} can range from 0 to 1, where 0 means complete sharing of genetic material and 1 means no sharing. The greater the genetic distance between populations, the less breeding there is between them and the more isolated they are from one another, in other words, values equal to 1, the populations do not share any allele with one another [2].

1.6.2.3 Polymorphism information content or power of information content

The polymorphism information content (PIC) indicates the polymorphic level of a given locus. It can be interpreted as the probability of being able to deduce which allele a parent has transmitted to the child Although, the PIC cannot be calculated when one parent is homozygous or when both parents and the child have the same heterozygous genotype [2].

1.6.2.4 Power of discrimination, power of exclusion and matching probability

Power of discrimination (PD) indicates the probability that, in a population, two randomly selected individuals, unrelated, will have different genotypes for a given locus. Thus, the higher the number of genetic markers analyzed higher the capability of discriminate individuals in a population [2,3].

Power of exclusion (PE) is the probability of a given locus to exclude one individual and can be define, in a typical paternity case, as the proportion of individuals that have a DNA profile different from that a randomly selected individual, permitting to exclude an alleged father in a population[58].

The matching probability (MP) is the probability that two randomly selected individuals, unrelated, will have identical genotype [2].

1.6.2.5 Paternity Index and Probability of paternity

The paternity index (PI) can be defined as how many times the person in test is more likely to be the biological father than a randomly selected individual. In the paternity calculations, the paternity index is the same as likelihood ratio. When is not a paternity in case, the likelihood ratio expresses the probability of finding this profile if the reference donor was the true contributor, compared to the probability of finding this profile if some other than the reference donor were the true contributor [2].

The probability of paternity (W) is the hypothesis that the man in test is the biological father of the child [2].

1.7 Portugal and the Portuguese population

Nowadays, Portugal is a continental country, with a total area of 92.072 km^2 and a current population of 10.493.530 (52% female, 48% male). It's located in Southwestern Europe (39 30 N, 8 00 W), on the Iberian Peninsula, is divided in eighteen districts and seven regions (Azores, Madeira, North, Center, Lisbon, Alentejo and Algarve) and has two archipelagos in the Atlantic Ocean (Madeira and Azores) (Figure 12). About 2.5% of the population living in Portugal is non Portuguese, being the major part people of African ancestry (mostly from Angola, Cape Verde and Mozambique) and more than 50% of the entire population resides in coastal areas, like Lisbon, Porto and Setubal [59–61].

Figure 12 - Portugal and the archipelagos of Madeira and Azores [60]**.**

Throughout the history, the territory known as Portugal, has been occupied by a succession of civilizations, like Celts, Phoenicians, Greeks, Romans, Carthaginians and Arabs. The Portuguese state was only instituted in 1139, when D. Afonso Henriques assumed the title of king of Portugal, and a lot of battles took place until 1297, when the *tratado de Alcanizes* was signed the country's borders to this day [62].

Also, due to its location, Portugal was a pioneer in marine exploration and expanded its territory in the fifteenth and sixteenth centuries, becoming a big empire with colonies in Africa, Asia, South America and Oceania. All this exploration and colonization have enabled the exchange of products and the movement of people between Portugal and those regions permitting the introduction of new habits and knowledge that influenced not only the countries culture but also its genetic diversity as well as the massive influx of refugees from former Portuguese colonies in Africa in the second half of the 1970s. Even though all those successive occupations and displacement of people across the world, the Portuguese population has been relatively homogeneous for most of its history and genetic statistics suggests a weak internal differentiation of the Portuguese population [62–64].

Others studies before have determined population data of chromosome STR loci in Portugal (studies performed in North, Center and South areas). Those previous investigations demonstrated that Hardy–Weinberg equilibrium is present in almost of all the STRs tested (STRs presents in commercial kits like Amp*Fl*STR® NGM™ STR loci; PowerPlex®16 System kit; PowerPlex® ES Monoplex System SE33; Amp*Fl*STR Profier Plus) and the forensic parameters indicated a high discriminating power of those loci. The results have been compared between the different Portugal areas (North, Centre and South) and no significantly variations have been found [52,53,65–68]. Also, the Portuguese allele frequencies distribution for the most recent five loci recommended by ENFSI and EDNAP groups, (D10S1248, D22S1045, D2S441, D1S1656, D12S391) have been compared to other European populations and it has been found that this distribution is analogous [53].

2. Objectives

The current markers used in routine casework in the SGBF-S laboratory of the Instituto Nacional de Medicina Legal e Ciências Forenses, delegação do Sul (INMLCF-S) are comprise of two commercial multiplex systems, Amp*Fl*STR® Identifiler Plus (Applied Biosystems) and PowerPlex® 16 HS (Promega Corporation), which amplifies a total of 18 STRs. In the major part of paternity test the combination used by the two kits is enough to obtain the discriminated power needed, although there are some complex cases where there is a need for more genetic information to obtain a strong enough association. Also it is more and more common to have casework involving trace or degraded DNA samples and that fact made necessary to implement new kits with more markers which, would be amplified in smaller amplicons and will allow obtaining a profile. Besides that, and like mentioned above, the implementations of some of these markers in routine casework are one of the recommendations of the European Network of Forensic Science Institutes DNA working group and the European DNA Profiling group to facilitate the data sharing across countries.

Therefore, the present study strived to contribute towards knowledge of the allele frequencies of the STR loci present in the commercial kit GlobalFiler™ Express in the South Portuguese population. The evaluation of sensibility, specificity, robustness and power of discrimination of those new markers, will permits the use of this new commercial kit in forensic identification, especially in complex cases like degraded samples or complex kinship tests. Specifically it's intended to:

- Obtain allele frequencies for each marker in South Portuguese population;
- Testing the efficacy of these markers for obtain relevant forensic statistic parameters;
- Perform implementation and internal validation of this new commercial kit, to be introduced in the laboratory's routine hereafter.

3. Material and Methods

3.1 Samples

In this population study were used 404 bloodstain reference samples of INMLCF-S, obtained under informed consent from unrelated individuals involved in paternity tests. 214 of the individuals were males and 190 were females.

All the individuals and their parents were natives of Portugal, residing in South region. The term South Portugal or region, in this case, refers to the region covered by the South delegation of INMLCF, namely, all the districts that are the line from Lisbon down (orange area in Figure 12).

3.2 Amplification and Typing

3.2.1 Polymerase Chain Reaction

DNA was amplified using the GlobalFiler™ Express Kit (Life Technologies) in a thermal cycler GeneAmp 9700 PCR system (Applied Biosystems), according to manufacturer's recommendations [50], but reducing the PCR final volume to half of the recommended quantity. This reduction is one of the steps performed in the internal validation and will be explained in more detail later.

In a laminar flow chamber at the extraction room, previously decontaminated 15 minutes with Ultraviolet light, was prepared the MicroAmp® optical 96-well reaction plate with the blood samples of each individual (using a micro puncher with 1.0 mm) and the positive and negative controls, according to the next table:

Table 2 - Required volume of Prep-n-Go™ buffer of the different types of samples

Type of Sample	Add
Test Samples	1.5 µL of Prep-n-Go™ Buffer 1.0 mm sample disc
Positive control	0 µL of Prep-n-Go™ Buffer 2 µL of DNA Control 007
Negative control	1.5 µL of Prep-n-Go™ Buffer 1.0 mm blank disc

After that first step, at a pre-PCR room, previously decontaminated 15 minutes with Ultraviolet light too, and in a laminar flow chamber, a mix was prepared into a polypropylene tube with the necessary volume of each component of the GlobalFiler™ Express kit in concordance with Table 3:

Table 3 - Required volume of each component of the GlobalFiler™ Express kit for one reaction.

Reagent	Volume required per sample
Master Mix	3.0 µL
Primer Set	3.0 µL

This mixture was vigorously mixed and briefly centrifuged and then vigorously mixed. In each reaction well were dispensed 6 µL. The plate was sealed with MicroAmp® clear adhesive film and centrifuged at 3000 rpm for approximately 20 seconds in a tabletop centrifuge. After that, the plate was put in a thermo cycler and the PCR was performed according to the following conditions (Table 4):

Table 4 - PCR conditions

Initial incubation step	22 Cycles		Final extension	Final hold
	Denature	Anneal/Extend		
95°C	94°C	60°C	60°C	4°C
1 min	3 sec	30 sec	8 min	∞

3.2.2 Capillary electrophoresis

The amplified products was separated and detected by capillary electrophoresis in an Applied Biosystems 3130XL Genetic analyser.

In a post-PCR room, a total of volume of Hi-Di™ Formamide and GeneScan™ 600 LIZ® v 2.0 Size Standard was dispensed to a polypropylene tube according to the quantities presents in Table 5 .

Table 5 - Required volume of each component needed to perform the capillary electrophoresis by sample.

Reagent	Volume per reaction
GeneScan™ 600 LIZ® Size Standard v2.0	0.2 µL
Hi-Di™ Formamide	9.8 µL

Subsequently, the mixture was vigorously mixed, briefly centrifuged and pipetted 10µL of the mixture formamide/size standard to each well of a MicroAmp® Optical 96-Well Reaction Plate and then was added 1µL of PCR product or 1µL of Allelic Ladder (one allelic ladder for run, in other words, one allelic ladder per 15 samples).

After the plate was prepared, it was sealed, briefly centrifuged and placed in a thermal cycler at 95ºC for 3 minutes and then placed another 3 minutes on ice. Finished those 3 minutes the plate was assembled on the autosampler to start the capillary electrophoreses.

3.3 Internal Validation Studies

In this internal validation study, the Quality Manual (internal document, not published) guidelines and parameters were followed. This document's chapter was elaborated in accordance with the SWGDAM recommendations and was development to facilitate the access to all the information needed to perform a internal validation and subsequent implementation of a new method.

3.3.1 Minimum Threshold Calculation

To determine the minimum threshold, it was prepared 5 mixtures that were used to evaluate baseline noise. Those 5 mixtures contained only size standard and formamide (0.2µL and 9.8µL, respectively) and were run 5 times each in two different runs in different days.

3.3.2 PCR Amplification optimization

Aiming to determine the appropriate PCR cycle number in order to obtain an optimized PCR reaction, a cycle sensitivity study was conducted: 12 samples were amplified at 27, 25, 24, 23 and 22 cycles following the manufacture's recommended amplifications conditions and the fragment detection and data analysis were performed using GeneMapper® ID-X 1.4 Software, according to the fabrication's recommendations.

3.3.3 Species specificity determination

In some forensic casework samples it may be present nonhuman DNA. In order to test the ability to detect and characterized unequivocally human DNA, a total of 27 animal samples from seven different species (nine horses; eight pigs; four dogs; one cat; one sheep; two donkeys and two cow), were amplified and genotyped using standard PCR and capillary electrophoresis conditions.

These set of animal samples are commonly used in SGBF-S to perform species specificity tests and so, DNA was already extracted with Chelex®.

3.3.4 Reproducibility study

To perform this study, 20 samples were typed according to the experimental proceeding by a collaborator different from the one that determined the initial profile. The obtained profiles were compared to the previously determined to verify if the results are the same.

3.3.5 Volume Reduction (Repeatability Test)

In this lab, in order to economic maximization of the commercial kit, the normal final volume was reduced to half. To test if this volume reduction was practicable, 20 samples were amplified with the volume recommended in the user guide (15µL) and with half volume (7.5µL) and manufactures' PCR conditions was performed The PCR products were detected using standard capillary electrophoresis conditions.

Obtained profiles were compared to evaluate if there was discordance in results. This test also permits to evaluate the repeatability of the method, once it evaluates the precision of two different measures in identical conditions.

3.3.6 Contamination Study

To analyze this characteristic, two mixtures were prepared, from four different individuals (two males and two females) with DNA previously extracted with Chelex®.

The two mixtures were prepared across the ranges: 1:1, 1:2, 1:5, 1:10 and 1:20. In the first mixture, the initial concentration were 1.2 ng/µL (female) 1.1 ng/µL (male) and female and male samples were mixed at deceased quantities of female DNA to constant quantity of male DNA, as Table 6 shows. The second mixture has samples with 2.8ng/µL (female) and 2.6ng/µL (male) of initial concentration and were prepared with a constant quantity of female DNA with decreased quantities of male DNA (Table 7).

Table 6 - Contamination test 1: mixtures design

Ratio	µL of Male sample	µL of Female sample
1:1	5	5
1:2	5	2.5
1:5	5	1
1:10	5	0.5
1:20	5	0.25

Table 7 - Contamination test 2: mixtures design.

Ratio	µL of Male sample	µL of Female sample
1:1	5	5
1:2	2.5	5
1:5	1	5
1:10	0.5	5
1:20	0.25	5

The PCR amplification was made using 1 µL of each mixture, according to the technical features as described previously. After the PCR amplification step, samples from contamination test were randomly put in another MicroAmp® Optical 96-Well Reaction Plate and loaded into the 3130 XL Genetic Analyser for fragment detection and data analysis using GeneMapper® ID-X 1.4 Software, with the threshold detection set at 50 RFU (relative fluorescence units).

3.3.7 Sensitivity study

In order to find out the optimal concentration of DNA input to obtain the best results, DNA control 007 was used to performed serial dilutions ranging from 2 ng/µl (not diluted) to 0.031 ng/µl (Table 8). The PCR amplification was made according to the technical features. Subsequently, a plate was prepared for fragment detection and data analysis using GeneMapper® ID-X 1.4 Software, using same conditions stated before.

Table 8 - Serial of dilutions performed to determine the optimal concentration of input template DNA.

Dilution	DNA volume (µL)	Water volume (µL)	Final concentration (ng/µL)
A	1 (original)	-	2
B	5 of A	10	1
C	5 of B	10	0.50
D	5 of C	10	0.25
E	5 of D	10	0.125
F	5 of E	10	0.062
G	5 of F	10	0.031

3.3.8 Concordance Study

For this purpose, 200 unrelated individuals living in Portugal, previously genotyped with Amp*Fl*STR® Identifiler Plus (Applied Biosystems) and PowerPlex® 16 HS (Promega Corporation) kits were genotyped with GlobalFiler™ Express kit according to the protocol already described.

The obtained profiles were compared to previously determined ones in order to evaluate if there were discordant result for the samples when using different kits with different performances and primer designs.

3.4 Population Study

All the allele identification and genotyping was performed used the GeneMapper® ID-X 1.4 Software (Applied Biosystems), using size comparison to the corresponding allelic ladder. All the multiplex kits panels, bins and analysis methods were obtained from the respective manufacturer (Applied Biosystems).

3.4.1 Populations parameters

Population parameters were calculated using the Arlequin v3.5 software [69]. Allele frequency was estimated for each marker as Ho, He, HWE and F_{ST}.

3.4.1.1 Observed Heterozigoty, Expected heterozigoty and Hardy–Weinberg equilibrium

The Ho is the observed quantity of heterozygotes, averaged over loci and the He is calculated as 1 minus the sum of the squared gene frequencies (the He is the value that would be obtain if the population is in HWE). The HWE can be illustrated mathematically by Equation 1 , where p and q represents the frequencies of alleles and p+q are always equal to one [3]**.**

Equation 1 - Formula to calculate HWE

$$p^2+2pq + q^2 = 1$$

3.4.1.3 Population pairwise genetic distances

Here, our South Portuguese population will be compared to North and Central Portuguese, as well as with other populations such as Korean, Spanish, Sicilian, Dutch, US Caucasian, Austrian, Lebanon and Swedish.

3.4.2 Forensic parameters

After the populations parameters are study, it became possible determine forensic parameters like, PIC, PD, PE, MP, PI and W. For this calculus were used PowerStats v1.2 software (Promega Corporation) [70].

3.4.2.1 Polymorphism information content or power of information content

PIC can be calculated by the equation below (Equation 2), when *pi* is the frequency of each distinct allele, and *n* is the number of distinct allele [2]. Although, the PIC cannot be calculated when one parent is homozygous or when both parents and the child have the same heterozygous genotype.

Equation 2 - Formula to calculate polymorphic information content

$$PIC = 1 - \sum_{i=1}^{n} {p_i}^2 - \left(\sum_{i=1}^{n} {p_i}^2\right)^2 + \pm \sum_{i=1}^{n} {p_i}^4$$

3.4.2.2 Power of discrimination

Equation 3 presents the formula to calculate the PD and in Equation 4 is the formula to combined power of discrimination, where, PD is the power of discrimination of a single locus, PM is the match probability of a single locus, PDcomb is the power of discrimination of numerous loci, PDi is the individual locus power of discrimination [2,3]

Equation 3 - Formula to calculate power of discrimination

$$PD = 1 - PM$$

Equation 4 - Formula to calculate combined power of discrimination

$$PDcomb = 1 - \prod_{i=1}^{n}(1 - PDi)$$

3.4.2.3 Power of exclusion

The formulas to calculate PE and combined power of exclusion are present in Equation 5 and Equation 6, respectively. Where, h is the heterozygosity, H is the homozygosity at the locus, L is the number of the loci, PEl is the exclusion probability for the lth locus and the π sign stands for multiplication [2].

Equation 5 - Formula to calculate the power of exclusion

$$PE = h^2(1 - 2hH^2)$$

Equation 6 - Formula to calculate combined power of exclusion

$$PEcomb = 1 - \prod_{l=1}^{L}(1 - PE_l)$$

3.4.2.4 Matching probability

MP is calculated by the formula below, in Equation 7, where, PM is the match probability, Pk the frequency of each distinct genotype and m is the number of distinctive genotypes. To calculate the combined probability of match with more than one locus it's made the product of the value for all the loci [2].

Equation 7 - Formula to calculate the matching probability

$$PM = \sum_{k=1}^{m} Pk^2$$

3.4.2.5 Paternity Index

The PI can be calculated using Equation 8, were H is the frequency of homozygotes. The paternity index of more than one locus, when they are inherited independently, is the product of all of the individual Paternity indexes [58].

Equation 8 - Formula to calculate the paternity index

$$PI_{typical} = \frac{1}{2H}$$

3.4.2.6 Probability of paternity

The probability of paternity is based upon Baye's theorem, and to calculate this parameter it is necessary the prior probability of the tested man is the father. The prior probability is typically 0.5 bases that this is a neutral, unbiased value: the present man in test is either the true father or he is not [71]. Thus, with this 0.5 prior value, the probability of paternity can be calculated by the next equation (Equation 9), where combined PI is the product of all the individual paternity indexes.

Equation 9 - Formula to calculate the probability of paternity

$$W = \frac{Combined\ PI}{[Combined\ PI\ \times 0.5 + (1 - 0.5)]}$$

4 Results and Discussion

4.1 Minimum Threshold Calculation

Minimum threshold is defined as the RFU value that separates a true DNA peak signal from the background noise [72]. The typical value applied currently in forensic routine, for all type of samples, ranges from 30 – 50 RFUs [30].

The background noise for all the 5 mixtures (Blanc controls) analysed were typically below 45 RFUs, only the red and purple channel have some peaks above this threshold (some of them were near or little above 50 RFUs). In the real case samples, independently of the channel, the negative controls have a baseline level always below 30 RFUs (Figure 13).

Based on evaluation of background noise and real case samples, values of 100 RFUs were chosen for reference samples and 50 RFUs for casework samples with low quantity of DNA or mixtures samples. The higher values for references samples were due to the height of the real peaks were there was no need to reduce more the threshold. However, it was verified that when a mixture sample was evaluated sometimes real peaks below the 50 RFUs were observed. Due to this fact, in those cases, a more carefully analysis was needed and the minimum threshold individually evaluated.

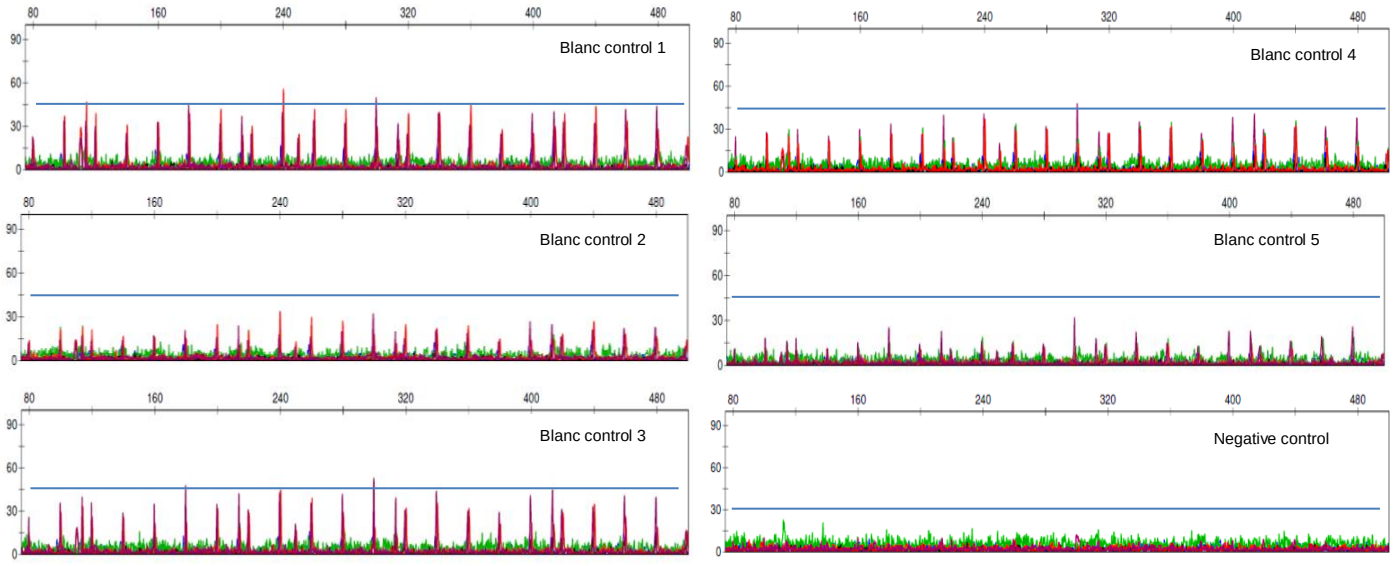


Figure 13 - Background noise evaluation: Electropherogram, with combined dyes, from the 5 blanc controls and one negative control

4.2 PCR Amplification optimization

According to the manufacturer user guide, when using Applied Biosystems 3130 XL Genetic Analyser system, PCR should result in profiles with heterozygous peaks ranging from 1000 to 3000 RFUs, without the occurrence of allelic drop out and with no minimal detected off-scale peaks [50].

To determine the ideal amplification cycle number, 12 samples and positive and negative controls were tested with 27, 25, 24, 23 and 22 amplifications cycles on a thermal cycler GeneAmp 9700 PCR system (Applied Biosystems).

A full profile was generated for all PCR cycle number tested, although with cycles ranging from 24 to 27 cycles a great number of pull-ups and off-scale data phenomena was observed, leading to difficulties on results interpretation. With a 23 cycles PCR, lesser or even none pull-ups and off-scale data problems were verified. However, in these conditions, a preferential amplification of smallest markers was detected. For this reason, amplification was reduced to a 22 cycle PCR and, in these conditions, a full profile was generated without any of the aforementioned problems (Figure 14).

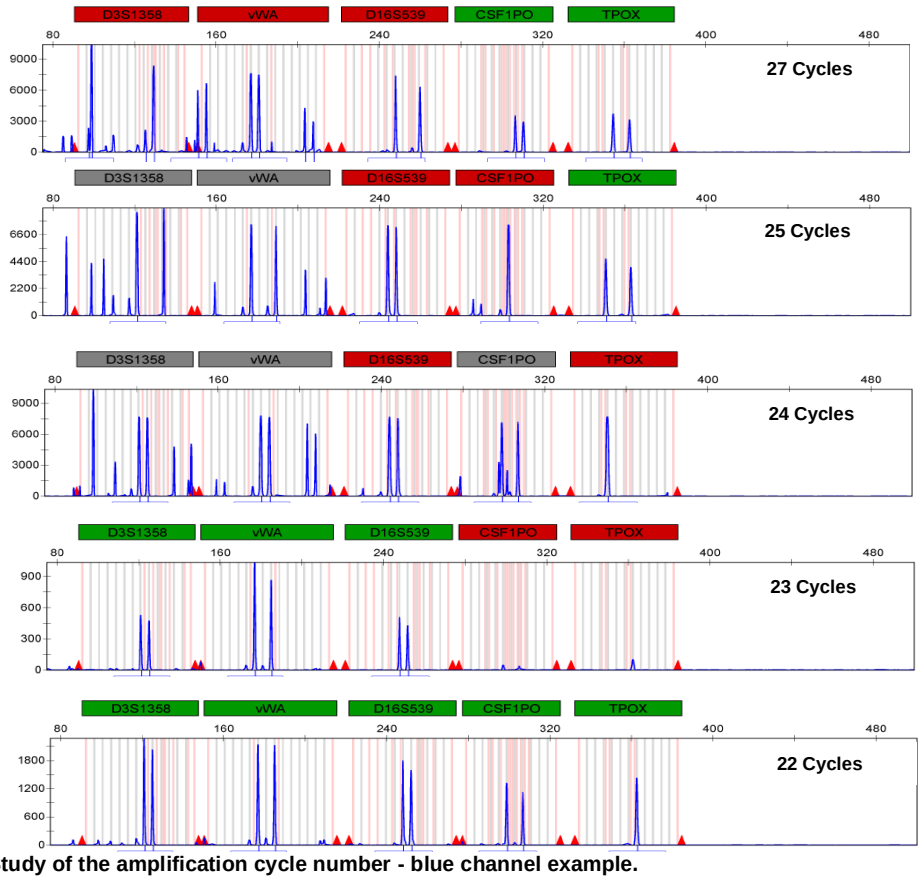


Figure 14 - Study of the amplification cycle number - blue channel example.
27 to 24 - it can be seen a lot off-scale data and more than 2 peaks per locus (from spectral pull-ups); 23 cycles - preferential amplification; 22 cycles – complete profile

4.3 Species specificity determination

For species specificity determination, 27 DNA samples from different animal species (nine horses; eight pigs; four dogs; one cat; one sheep; two donkeys and two cows) were tested. Unfortunately, primates samples were not available, making not possible to study the specificity of the molecular markers using this animal group, which is closely related to humans.

The analyses of all animal samples presented negative (Figure 15), i.e., it was not detected any amplification peaks, and so this test confirmed that GlobalFiler^{TM} Express provides the required specificity for human identification, being highly specific for human DNA.

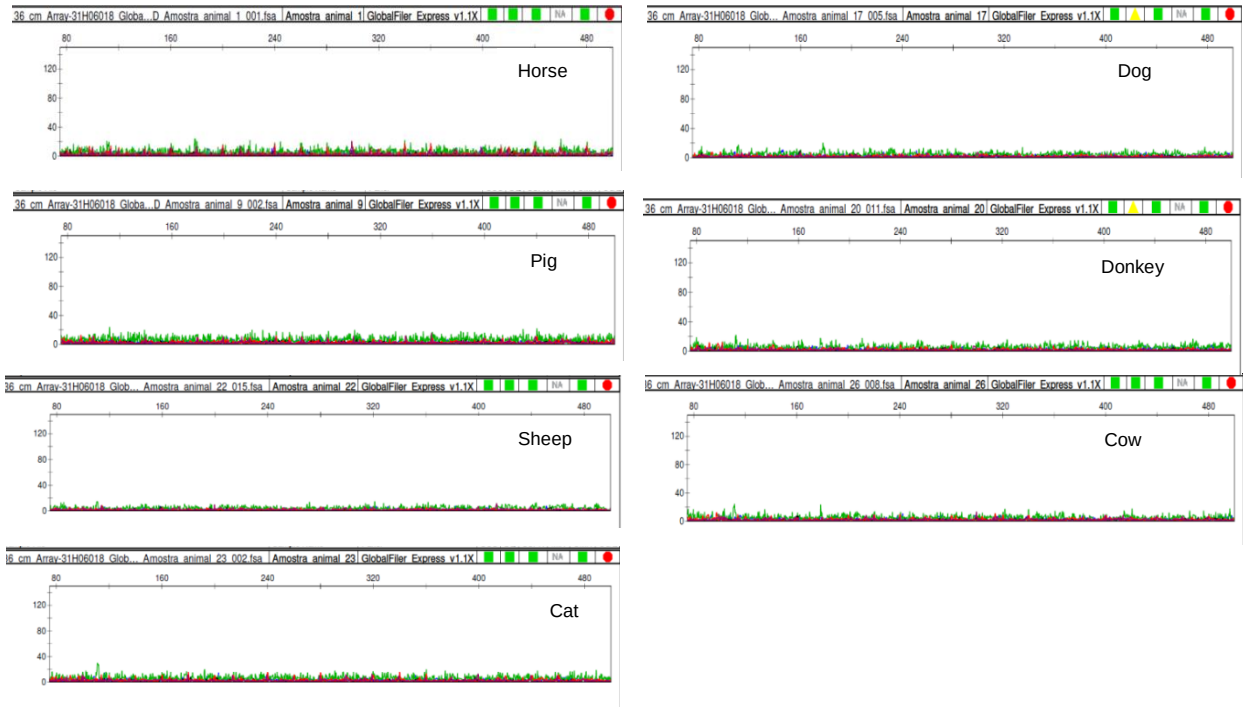


Figure 15 - Electropherograms (with combined dyes) from species specificity test

4.4 Reproducibility study

The reproducibility study was performed by two different collaborators. Each one genotyped the same 20 known samples, according to the experimental proceeding.

Obtained profiles were compared and no discordance data was found. This test permitted to conclude that the method is reproducible, confirming that sample genotyping using this procedure, by different technicians, will give the same profile.

4.5 Repeatability test

In order to increase the economic maximization of the commercial kit, an half of the recommended PCR mix volume assay was tested. To determine if this reduction does not affect the results, 20 samples were genotyped using the manufacturer volume protocol of 15µL and the half of the volume essay, 7.5µL. The obtained profiles were compared and the results obtained with half of the volume procedure were equal to the ones obtained with the total volume protocol.

It may be concluded that the reduction of total volume is practicable and that this method presents repeatability, as the same samples in different conditions present the same profiles.

4.6 Contamination Study

In these experiments, the ability of detecting mixtures in samples was tested by the examination of the appearance of more than two peaks per locus.

The results obtained from the two mixtures in different proportions (1:1; 1:2; 1:5; 1:10 and 1:20) were evaluated with the threshold for detection set at 50 RFU. In all the mixtures it was possible to detect clearly that the samples had more than one contributor, by the presence in the profile of three or more alleles per locus (Figure 16). This is a very important factor since an accidental contamination can occur, and the method need to be able to discriminated clearly single source samples from mixture or contaminated ones.

It also can be observed that, even when the minor contributor is present in a 1:20 ratio, the contamination is easily detected. However, some of the minor contributor peaks are visualized but not detect by GeneMapper software as alleles due to the 50 RFUs threshold (some examples marked with an arrow on Figure 17).

Thus, it is extremely important that in the presence of mixture samples, when a minor contributor may be present, a more careful analysis must be performed, following the current guidelines for mixture interpretation in order to prevent loss of valuable information [73,74].

Figure 16 - Electropherograms from contamination study: Ratio 1:1.The presence of three or more allele per locus proves the capability of the method to distinguish mixture samples.

Figure 17 - Electropherograms from contamination study: Ratio 1:20. The presence of three or more allele per locus prove the capability of the method to distinguish mixture samples, even in this proportion, although some real peaks are below the 50 RFU threshold and are not marked as an allele (some examples marked with an arrow)

4.7 Sensitivity Study

When DNA concentration is lower than the optimal concentration, stochastic effects, such as heterozygous imbalance or allelic drop-in and drop-out, can alter the genetic profile In the other hand, when the quantity of DNA input is too much, different artifacts can be detected, such as pull-up and/or stutter effect. All these phenomena difficult and compromise profile interpretation [3,75].

To determine the optimal DNA input in order to obtain correct and reproducible results and the limit of detection of the system (the point where the software do not detect the presence of a real allele) a serial of dilutions were performed ranging from 2 ng/µL to 0.031 ng/µL. Complete profiles were obtained in samples ranging from 2 to 1 ng/µL (Figure 18 and Figure 19), and incomplete profiles were detected up to 0.25 ng/µL (Figure 20). When final concentration was below 0.25 ng/µL, the limit of detection of the system was reached and it was not detected any peak (Figure 21). The achieved sensitivity values were not very high, since this commercial kit was developed to be used in direct amplification of reference samples, in which there are no problems of input quantity of DNA. Usually, the sensitivity range was not determined in development studies performed by the manufacturer, although it was determined for the GlobalFiler™ which is designed for casework samples and, as expected, this kit presents a low sensitivity when compared to the GlobalFiler™ kit [50,75].

Figure 18 - Sensitivity Test – Electropherograms with 2ng/µ (complete profile).

Figure 19 - Sensitivity Test – Electropherograms with 1 ng/µL (complete profile).

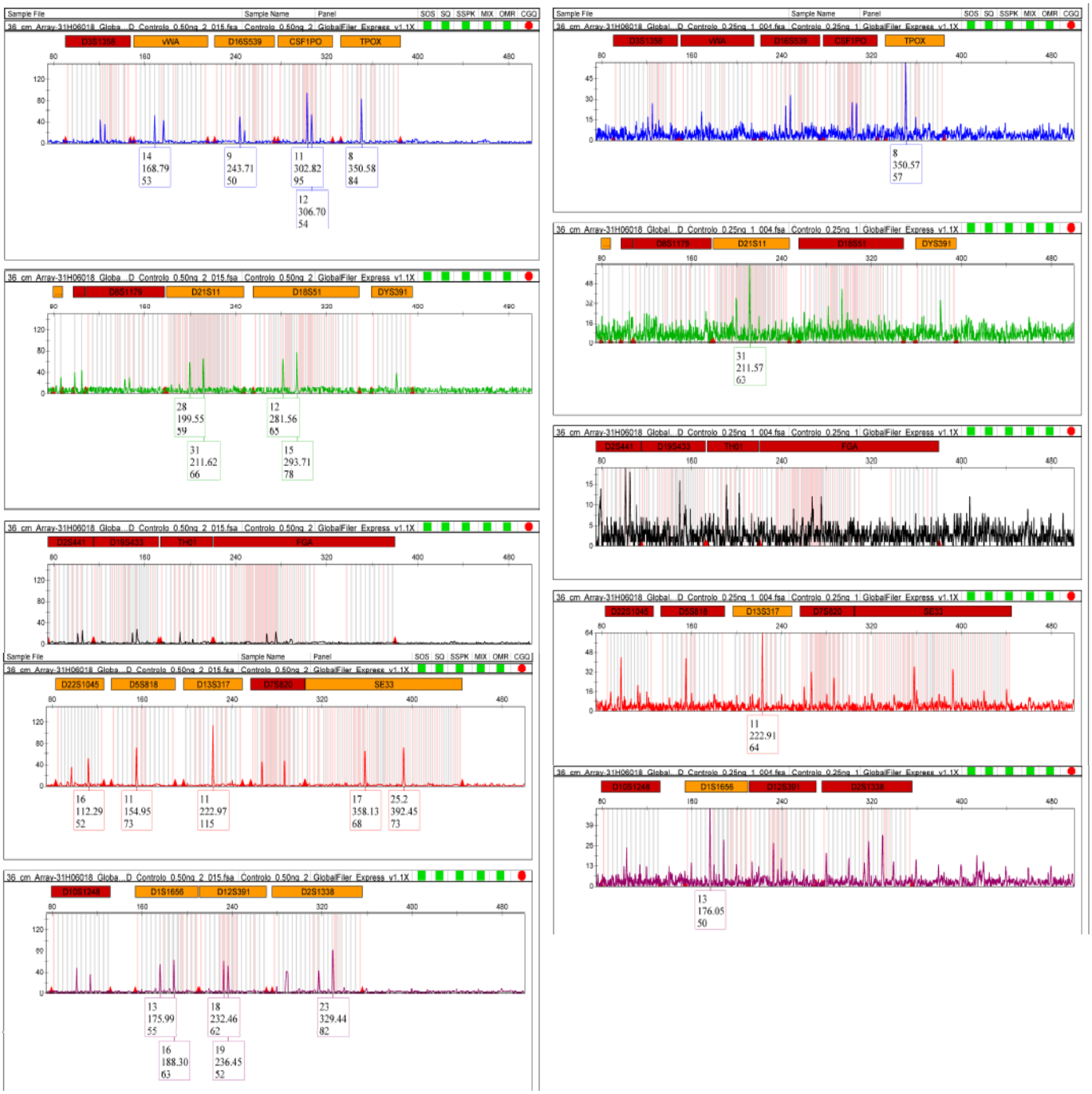


Figure 20 - Sensitivity Test – Electropherograms with 0.5ng/µL and 0.25ng/µL (incomplete profiles).

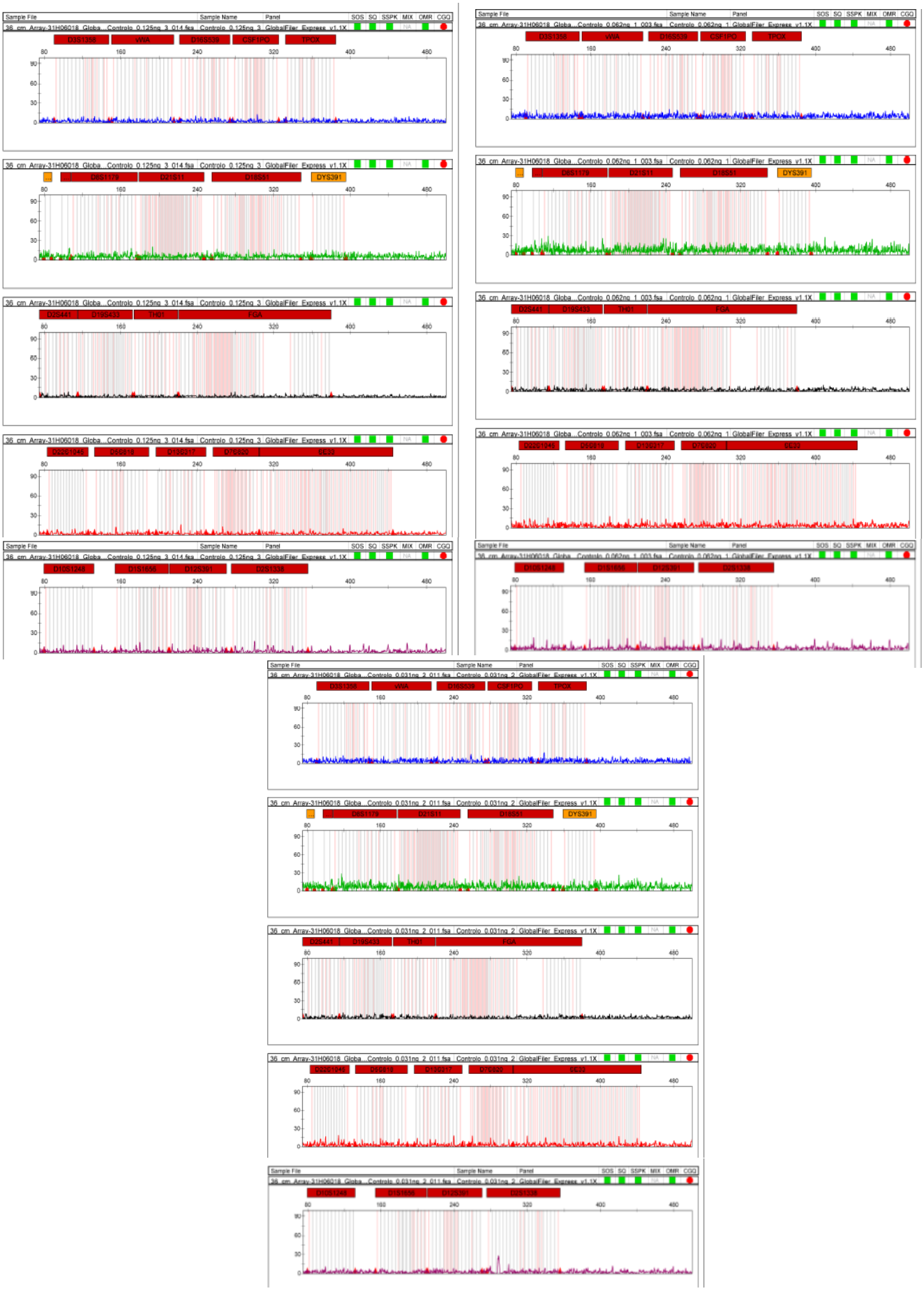


Figure 21 - Sensitivity Test – Electropherograms with 0.125ng/µL, 0.0625ng/µL and 0.031ng/µL (profiles without any amplified peak).

4.8 Concordance Study

This study permits the observation of any discordant genotype results for the same samples genotyping, using different commercial kits. Discrepancies between multiplex kits may be caused by the presence of microvariations outside the repeat motifs or due to silent alleles [76]. Once the two kits used in routine share 16 loci with GlobalFiler™ Express, it was evaluated the possibility of allelic drop-out or null allele be present in the data set.

The obtained results revealed 100% concordance between the typing results with GolbalFiler™ Express and the two other kits (Amp*Fl*STR® Identifiler Plus (Applied Biosystems) and PowerPlex® 16 HS (Promega Corporation)) used in routine casework.

4.9 Population Study

In this study, a total of 404 samples from Portugal unrelated individuals living in South Portugal area and involved in paternity testing were genotyped with GlobalFiler™ Express amplification kit. Population parameters were calculated using Arlequin V3.5 software and PowerStats V1.2.

4.9.1 Population Parameters

Rare alleles were observed in two different markers in different individuals. The variant 18.1 STR was found in the SE33 marker in a male individual and the variant 17.1 was found in D1S1656 in a female individual. Until then, these alleles were only reported in criminal cases, and with no known suspects, making these the two first known individuals were they were found. Variant 18.1 in SE33 was reported by the Police of the Government of Catalonia-Mossos d'Esquadra (Spain) [77] and discovery of 17.1 in D1S1656 loci was reported by the Forensic Genetics Department, Autonomous Police of the Basque Country [78].

The observed allele frequencies distribution for the 21 genetic autosomal markers, present in the commercial kit GlobalFiler™ Express, were calculated using Arlequin V3.5 software and are listed in Table 9.

Table 9 - Allele frequencies for the 21 autosomal markers in South Portugal

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01
3											0.0012
6					0.0074						0.1918
7				0.0012	0.0037						0.1782
8			0.0149	0.0074	0.5396	0.0149			0.0012		0.1139
8.3											0.0012
9			0.1250	0.0210	0.1126	0.0050			0.0012		0.2203
9.3											0.2748
10			0.0693	0.2847	0.0619	0.0866		0.0124	0.1918	0.0012	0.0186
10.2								0.0012			
11			0.2871	0.3007	0.2500	0.1126		0.0099	0.3119	0.0025	
11.2											
11.3									0.0829		
12	0.0025		0.3032	0.3205	0.0235	0.1399		0.1262	0.0334	0.0978	
12.1											
12.2										0.0012	
13	0.0062	0.0062	0.1584	0.0495	0.0012	0.3094		0.1374	0.0347	0.2797	
13.2										0.0136	
14	0.0965	0.1040	0.0408	0.0136		0.1894		0.1448	0.2933	0.3156	
14.2										0.0347	
14.3										0.0012	
15	0.2921	0.1374	0.0012	0.0012		0.1163		0.1460	0.0446	0.1460	
15.2										0.0396	
15.3											
16	0.2488	0.2352				0.0248		0.1448	0.0037	0.0458	
16.2										0.0161	
16.3											
17	0.2141	0.2661				0.0012		0.1374	0.0012	0.0050	
17.1											
17.2											
17.3											
18	0.1287	0.1720						0.0668			
18.1											
18.3											
19	0.0111	0.0693						0.0359			
19.2											
19.3											
20		0.0099						0.0186			
20.2											
20.3											
21								0.0062			
21.2											
22								0.0087			
22.2											
23								0.0037			
23.2											
24											
24.2											
25							0.0050				
25.2											
26											
26.2											
27							0.0198				
27.2											
28							0.1510				
28.2											
29							0.2154				
29.2											
30							0.2537				
30.2							0.0396				
31							0.0458				
31.2							0.1102				
32							0.0074				
32.2							0.1163				
33											
33.2							0.0297				
34											
34.2							0.0062				
35											
36											

Table 9 (continued)

Allele	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D1S1656	D12S391	D2S1338
3										
6				0.0025						
7					0.0173	0.0012				
8			0.0062	0.1324	0.1411					
8.3					0.0012					
9			0.0347	0.0594	0.1448					
9.3										
10		0.0012	0.0606	0.0545	0.2488			0.0025		
10.2										
11		0.1040	0.3465	0.3305	0.2451		0.0037	0.0594		
11.2						0.0012				
11.3										
12		0.0111	0.3602	0.2562	0.1733	0.0025	0.0557	0.1522		
12.1			0.0012							
12.2										
13		0.0111	0.1856	0.1163	0.0198	0.0173	0.2760	0.0483		
13.2						0.0025				
14		0.0421	0.0050	0.0458	0.0074	0.0347	0.2958	0.0990		0.0012
14.2										
14.3								0.0025		
15		0.3663		0.0025	0.0012	0.0569	0.2030	0.1460	0.0446	
15.2										
15.3								0.0656		
16		0.3552				0.0532	0.1361	0.1052	0.0334	0.0458
16.2						0.0025				
16.3								0.0668		
17	0.0025	0.0978				0.0569	0.0297	0.0545	0.1027	0.2599
17.1								0.0012		
17.2						0.0012				
17.3								0.1349	0.0223	
18	0.0074	0.0111				0.0829		0.0037	0.1881	0.0681
18.1						0.0012				
18.3								0.0446	0.0248	
19	0.0792					0.0767		0.0025	0.1126	0.1238
19.2	0.0012					0.0037				
19.3								0.0099	0.0186	
20	0.1262					0.0483			0.1176	0.1411
20.2						0.0099				
20.3								0.0012	0.0012	
21	0.1807					0.0260			0.1040	0.0421
21.2	0.0050					0.0161				
22	0.1819					0.0087			0.1077	0.0334
22.2	0.0050					0.0384				
23	0.1547					0.0012			0.0755	0.1064
23.2	0.0012					0.0384				
24	0.1312								0.0260	0.0891
24.2						0.0421				
25	0.0718								0.0186	0.0767
25.2						0.0408				
26	0.0433								0.0025	0.0111
26.2						0.0396				
27	0.0074									0.0012
27.2						0.0606				
28	0.0012									
28.2						0.0730				
29						0.0012				
29.2						0.0705				
30										
30.2						0.0446				
31										
31.2						0.0272				
32										
32.2						0.0062				
33						0.0025				
33.2						0.0025				
34						0.0025				
34.2						0.0025				
35						0.0012				
36						0.0012				

Beside the 21 autosomal markers, the kit genotypes 2 markers located in the Y chromosome. The frequencies for those two markers were calculated using PowerStats V 1.2 software.

The Y InDel, for all the samples presented the ancestral insertion form, due to the fact that individuals only presents the deletion form if they have Asian origin [54], which were not present in our sampling.

The DYS391 Y-STR frequencies are present in Figure 22. This locus is not much polymorphic; the most frequent alleles were 10 and 11, with 0.458 and 0.463 frequencies, respectively. That result are in concordance to the other studies performed in Portugal with this marker [79,80] and with studies for other Europeans populations, like Southwestern Poland [81] and Southern Italian populations [82], were those two alleles were also the most common ones.

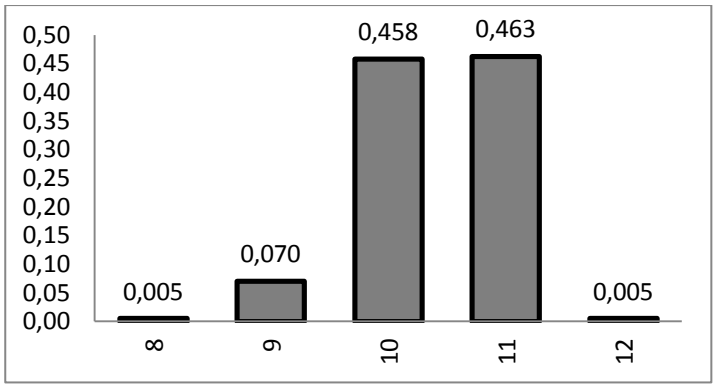


Figure 22 - Graphic representation of DYS391 allele frequencies.

Since these two markers are located on a sexual chromosome, they are dependent of sex and are transmitted only from father to son without suffering recombination. For this reason, they are not included in the other future forensic and population parameters determination.

Allele distribution p-values results confirmed that all autosomal loci were in Hardy-Weinberg equilibrium, applying a 0.05 significance level (table 10).

Table 10 - Hardy-Weinberg equilibrium.

Locus	Observed Heterozygosity	Expected Heterozygosity	p-value	standard deviation
D3S1358	0.7822	0.7819	0.5495	0.0003
vWA	0.8045	0.8107	0.1316	0.0003
D16S539	0.7723	0.7792	0.8098	0.0003
CSF1PO	0.6757	0.7235	0.5126	0.0005
TPOX	0.6312	0.6300	0.8957	0.0002
D8S1179	0.8193	0.8153	0.6125	0.0005
D21S11	0.8564	0.8368	0.2409	0.0003
D18S51	0.8861	0.8777	0.4557	0.0004
D2S441	0.7871	0.7697	0.7646	0.0004
D19S433	0.8119	0.7869	0.5966	0.0004
TH01	0.7822	0.7951	0.5741	0.0005
FGA	0.8861	0.8648	0.9000	0.0003
D22S1045	0.6931	0.7180	0.0719	0.0002
D5S818	0.7104	0.7117	0.3932	0.0004
D13S317	0.8267	0.7865	0.1053	0.0003
D7S820	0.7921	0.8074	0.1385	0.0002
SE33	0.9406	0.9494	0.1198	0.0002
D10S1248	0.7748	0.7736	0.8711	0.0003
D1S1656	0.8936	0.8978	0.1635	0.0002
D12S391	0.9183	0.8950	0.5659	0.0004
D2S1338	0.8589	0.8634	0.0747	0.0002

These results are concordant to the previously studies performed in South Portuguese population with Amp*Fl*STR® NGM™ [53] and Amp*Fl*STR® Profiler Plus™ [83], where no Hardy–Weinberg deviation was found. Even thought, in the study performed by Cruz *et al*. (2006) a deviation from Hardy–Weinberg (HWE) was found in SE33 [84]. This deviation may occurred by the presence of 43 samples from African people that makes those sampling more heterogeneous and introduced rare alleles, like 9.2 and 39.2 that was not present in our South Portuguese population. Alleles are expected to remain unchanged from one generation to the next and so, rare alleles are not expected to be found in a population. Their presence leads to high variances, because they are always virtually in heterozigoty condition, causing deviation from HWE.

4.10 Forensic Parameters

All the forensic parameters were calculated using PowerStats V 1.2 software and Microsoft® Excel® 2010. Results are showed in Table 11 and Table 12.

Table 11 - Forensic parameters of interest.

Allele	MP	PD	PIC	PE	PI	Homo	Hetero
D3S1358	0.0846	0.9154	0.7465	0.5664	2.2955	0.2178	0.7822
vWA	0.0652	0.9348	0.7829	0.6073	2.5570	0.1955	0.8045
D16S539	0.0809	0.9191	0.7452	0.5486	2.1957	0.2277	0.7723
CSF1PO	0.1212	0.8788	0.6710	0.3917	1.5420	0.3243	0.6757
TPOX	0.1818	0.8182	0.5806	0.3300	1.3557	0.3688	0.6312
D8S1179	0.0592	0.9408	0.7910	0.6354	2.7671	0.1807	0.8193
D21S11	0.0489	0.9511	0.8159	0.7076	3.4828	0.1436	0.8564
D18S51	0.0303	0.9697	0.8637	0.7672	4.3913	0.1139	0.8861
D2S441	0.0944	0.9056	0.7335	0.5754	2.3488	0.2129	0.7871
D19S433	0.0766	0.9234	0.7567	0.6213	2.6579	0.1881	0.8119
TH01	0.0743	0.9257	0.7623	0.5664	2.2955	0.2178	0.7822
FGA	0.0360	0.9640	0.8485	0.7672	4.3913	0.1139	0.8861
D22S1045	0.1206	0.8794	0.6712	0.4176	1.6290	0.3069	0.6931
D5S818	0.1367	0.8633	0.6596	0.4445	1.7265	0.2896	0.7104
D13S317	0.0833	0.9167	0.7562	0.6496	2.8857	0.1733	0.8267
D7S820	0.0667	0.9333	0.7781	0.5844	2.4048	0.2079	0.7921
SE33	0.0077	0.9923	0.9457	0.8788	8.4167	0.0594	0.9406
D10S1248	0.0885	0.9115	0.7364	0.5531	2.2198	0.2252	0.7748
D1S1656	0.0226	0.9774	0.8877	0.7823	4.6977	0.1064	0.8936
D12S391	0.0231	0.9769	0.8846	0.8330	6.1212	0.0817	0.9183
D2S1338	0.0338	0.9662	0.8488	0.7125	3.5439	0.1411	0.8589

MP = Matching Probability; PD = Power of Discrimination; PIC = Polymorphic information content;
PE =Power of Exclusion; PI = Typical Paternity Index;
Homo = Frequency of homozygosity; Hetero = Frequency of heterozygosity

Table 12 - Interesting forensic parameters: combined results of the 21 markers.

Parameter	Value
Combined Matching Probability	1.8356×10^{-26}
Combined Power of Discrimination	0.99999999999999999999999981765
Combined Power of Exclusion	0.99999999966339800
Combined Paternity Index (likelihood ratio)	2.666.045.842
Typical Probability of paternity (W)	0.99999999962491300

As it can be observed in Table 11, the Matching probability (MP) is situated between 0.0077 and 0.1818. The highest value for this parameter is given by TPOX marker and the lowest by SE33. The combined Matching probability is $1.8356x10^{-26}$ (Table 12), meaning that two randomly selected individuals, unrelated, will have $1.8356x10^{-26}$ of probability of have identical genotypes.

Relatively to the power of discrimination (PD), the values range between 0.8182 and 0.9923, being the highest and the lowest values given, respectively, by SE33 and TPOX markers (Table 11). Thus, all present a power of discrimination above 0.80, that is one of the initial criteria's to select the STR loci used for forensic identifications [85]. The 21 autosomal markers together present a power of discrimination of 0.99999999999999999999999981765 (Table 12), meaning that in a population, two randomly selected individuals, unrelated, will have 0.99999999999999999999999981765 of probability of had different genotypes.

The most informative marker is SE33, which present the highest polymorphic information content (PIC) of 0.9457. All of the other markers have PIC values above 0.58 (Table 11).

Also, it is possible to observe in Table 11 that the power of exclusion (PE) is 0.3300 for TPOX marker (minimum value) and 0.8788 for SE33 (maximum value). Combined all the markers the power of exclusion is 0.99999999966339800 (Table 12). That means that a probability of this set of loci to exclude one individual is 0.99999999966339800.

Typical Paternity Index is situated between 1.3557 (TPOX) and 8.4167 (SE33) (Table 11). Being the combined typical paternity index equal to 2.666.045.842, in other words, with the 21 markers the individual in test is 2.666.045.842 billion times more likely to be the biological father than a randomly person and the typical probability of paternity, with a prior probability of paternity equal to 0.5, is 0.99999999962491300 (Table 12).

Homozygosity and heterozygosity parameters were also determined. The SE33 was the more heterozigotic marker (0.9406) while TPOX was the more homozigotic (0.3688) (Table 11).

In sum, among the 21 markers presents in the commercial kit GlobalFiler^{TM} Express the more polymorphic locus and consequently more informative and useful in forensic identification is SE33. This marker has the highest PIC, PD, PE, PI and the lowest MP and homozigoty frequency. These results are concordant to others studies performed anteriorly, that showed that TPOX is the molecular marker, present in CODIS, that revealed lowest variation among individuals [3].

Obtained results are according to the previously determined from other authors for South Portugal population, using others identification kits.. A study with SE33 confirmed the high degree of polymorphism of this locus [84] and similar results were also found in studies performed in North and Central Portugal population where SE33, when tested, were the most polymorphic locus and TPOX the least polymorphic one. Those studies confirmed that South Portuguese population was not significant different from North and Central Portugal area ones [68,86].

A comparison study was performed between South, North [52,68] and Central [86,87] Portugal populations, as well as, others populations, namely, Spanish [88], US Caucasian [89], Korean [90], Lebanese [91], Dutch [92], Austrian [93,94], Swedish [95–97] and Sicilian (Italy) [98,99]. Due to the fact that the kit contains a large number of loci and some of them were only recently added to ESS, it was difficult to find a population study that included all of them. Furtherin some populations in order to obtained data information for all loci it was necessary to gather data from 2 or 3 different studies. For the Portugal Central area the comparison was performed not considering SE33, because there is no published data.

The comparison study was made with an F-statistics test, applying a significance level of $p < 0.05$, where the null hypothesis (H0) states that the differences between populations are not statistically significant and the alternative hypothesis (H1) state that the observed differences are statistically significant. Thus, when the p-value is exceeded (i.e. $p > 0.05$) the H0 is accepted. The obtained results from the comparison study are present in Table 14 and Table 14.

Table 13 - Population pairwise F_{ST}

	Portugal South	Portugal North	Portugal Central	Austria	US Caucasian	Dutch	Sweeden	Spain	Sisilia (Italy)	Lebanon	Korea
Portugal South	0.00000										
Portugal North	-0.00165	0.00000									
Portugal Cental	0.00425	0.00643	0.00000								
Austria	-0.00173	-0.00251	0.00048	0.00000							
US Caucasian	-0.00184	-0.00198	0.00330	-0.00246	0.00000						
Dutch	0.00074	0.00123	0.00334	-0.00108	-0.00089	0.00000					
Sweeden	0.00153	0.00087	0.01093	0.00216	-0.00032	0.00121	0.00000				
Spain	0.00203	-0.00045	0.00307	-0.00256	-0.00068	-0.00034	0.00217	0.00000			
Sisilia (Italy)	0.00189	0.00080	0.00473	0.00007	-0.00043	0.00147	0.00256	0.00002	0.00000		
Lebanon	0.00420	0.00892	0.00218	0.00416	0.00475	0.00645	0.01296	0.00904	0.00499	0.00000	
Korea	0.01314	0.01297	0.02784	0.01763	0.01949	0.02558	0.02675	0.02848	0.02868	0.02592	0.00000

Table 14 - Comparison between South Portugal and other populations (including North and Central Portugal areas).
Table of probability values (p +/- standard deviation), with statistical differences (p < 0.05) marked in bold type.

	Portugal South	Portugal North	Portugal Cental	Austria	US Caucasian	Dutch	Sweeden	Spain	Sisilia (Italy)	Lebanon	Korea
Portugal South	*										
Portugal North	0.68468+-0.0433	*									
Portugal Cental	**0.01802+-0.0121**	**0.01802+-0.0121**	*								
Austria	0.75676+-0.0279	0.79279+-0.0183	0.30631+-0.0433	*							
US Caucasian	0.91892+-0.0228	0.71171+-0.0213	**0.01802+-0.0182**	0.90090+-0.0384	*						
Dutch	0.21622+-0.0388	0.19820+-0.0379	**0.00000+-0.0000**	0.69369+-0.0408	0.71171+-0.0345	*					
Sweeden	0.17117+-0.0316	0.24324+-0.0408	**0.00000+-0.0000**	0.15315+-0.0333	0.43243+-0.0265	0.05405+-0.0148	*				
Spain	0.14414+-0.0309	0.49550+-0.0412	**0.02703+-0.0139**	0.84685+-0.0365	0.53153+-0.0566	0.39640+-0.0474	0.11712+-0.0333	*			
Sisilia (Italy)	0.12613+-0.0388	0.31532+-0.0365	**0.00000+-0.0000**	0.48649+-0.0578	0.45045+-0.0650	0.09910+-0.0286	0.06306+-0.0194	0.38739+-0.0237	*		
Lebanon	**0.01802+-0.0121**	**0.00901+-0.0091**	**0.04505+-0.0152**	**0.04505+-0.0152**	**0.03604+-0.0278**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00901+-0.0091**	*	
Korea	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	*

The genetic distances between different populations are present in Table 13. Due to the fact that the F_{ST} is not only sensitive to the level of genetic differentiation among populations, but also to the allele frequency distribution, very low or even negative values appears reflecting the fact that more variance exists within than across population. This negative values, as recommended, will be interpreted as zero [100,101]. Therefore, as expected, the European populations are very similar and the most distant population is the one from Korea. South and Central Portugal area populations reveal an F_{St}=0.00425, which is an unexpected result since the three areas always appears as proximate, this result may be a consequence of fact that the central sampling are much bigger than the used in the present study (2125 and 404, respectively for Central and South area and 213 from North Portugal).

The three Portugal areas were compared and no significant deviation was found between North and South. Nevertheless the Central area, as noted before by the F_{St}, values revealed to be significantly different when compared with the other two areas (Table 14). However, after applying the Bonferroni correction for multiple tests ($P < 0.05 / 21 = 0.0024$), no significant deviations were considered between the three Portuguese areas.

The comparison of South Portuguese population with other ones revealed significant differences only with Lebanon and Korea (Table 14). However, after applying Bonferroni correction ($P < 0.0024$), no significant deviations should be considered between south Portuguese population and Lebanon.

The obtained results are concordant to other population studies, where no differences were observed among European populations [88,95,99].

The US Caucasian population [89] revealed no differences neither from south Portuguese population nor from other European populations tested. These results are concordant to the previously obtained in a study performed with the Swedish population, were the US Caucasian population only showed a significant difference in one of the fifteen tested loci (CSF1PO) [95].

Only few population studies with STRs for Lebanon population exists and even some do not perform populations comparison. However, Chouery and

collaborators (2010) compared Lebanese with Jordan, Egypt, Iraq, Turkey, Saudi Arabia, Syria and European Caucasian populations. They found out that the most different was the European Caucasian (with six markers significantly different) [102], and the results were in concordance with the ones obtained in this study.

A population study performed in Korea by Hong *and collaborators* (2012) obtained similar results to the one presented in this study: the Korean population demonstrated to have most genetic affinity with the Asian Group (Chinese, Thai, Japanese, among others) and no relation with European populations (Russian, Spanish, Italian, Portuguese, etc.) [90]. As well as a study performed with the Italian population reveals that Korea, China, Malay and, India were significantly different from Spain and Italian data [103].

Analysis of Molecular Variance (AMOVA) was used to estimate population differentiation. Populations were subdivided by geographic location. In the first test the populations were clustered into 3 groups, European vs American vs Asian. The second test was performed with 4 groups Mediterranean Europe vs North of Europe vs American vs Asian. The next test the countries were clustered by Mediterranean Europe vs North of Europe vs American vs East Asian (Korea) vs Western Asia (Lebanon). After those first tests, and using the F_{ST} results previously determined as reference, it was grouped the US Caucasian population with the European countries to evaluate the differentiation between those populations.

The results are present in Table 15.

Table 15 - AMOVA test results:

Source of variation	Percentage of variation					
Test	A	B	C	D	E	F
Among groups	0.29	0.05	0.43	0.50	0.53	1.22
Among populations within groups	0.50	0.57	0.27	0.24	0.21	0.26
Within populations	99.21	99.38	99.30	99.26	99.26	98.53
FSC P-value	0.00499 0.00000+- 0.00000	0.00568 0.00000+- 0.00000	0.00269 0.00000+- 0.00000	0.00243 0.00000+- 0.00000	0.0213 0.00000+- 0.00000	0.00261 0.00000+- 0.00000
FST P-value	0.00788 0.00000+- 0.00000	0.00616 0.00000+- 0.00000	0.00696 0.00000+- 0.00000	0.00737 0.00000+- 0.00000	0.00740 0.00196+- 0.00136	0.01475 0.00000+- 0.00000
FCT P-value	0.00291 0.20137+- 0.01437	0.00048 0.35973+- 0.01563	0.00428 0.16422+- 0.01207	0.00496 0.10850+- 0.0101	0.00529 0.06354+- 0.00799	0.01217 0.08700+- 0.00828

Significance tests (1023 permutations)

A = Portugal (3 areas) +Spain + Italy + Austrian+ Swedish + Dutch vs US Caucasian vs Lebanon + Korea

B = Portugal (3 areas) +Spain + Italy vs Austrian+ Swedish + Dutch vs US Caucasian vs Lebanon + Korea.

C = Portugal (3 areas) +Spain + Italy vs Austrian+ Swedish + Dutch vs US Caucasian vs Lebanon vs Korea

D = Portugal (3 areas) +Spain + Italy + US Caucasian vs Austrian+ Swedish vs Lebanon vs Korea

E = Portugal (3 areas) +Spain + Italy vs Austrian+ Swedish + US Caucasian vs Lebanon vs Korea

F= Portugal (3 areas) +Spain + Italy + Austrian+ Swedish + Dutch +US Caucasian vs Lebanon vs Korea.

The AMOVA results, in all the tests, as expected, show that most of the genetic variation is within populations and not among populations. Among groups the major diversity obtained was 1.22%. This value was observed when tested Lebanon vs Korea vs all European countries with the US Caucasian population. It seems to be the best separation since it was observed also the minor variation within populations (98.53%) and one of the minor values of variation among populations within groups (0.26%).

The most percentage of variation within populations (99.38%) and minor among groups (0.05%) were observed when was tested the Mediterranean

Europe vs North of Europe vs American vs Asian. These values indicated an enormous similarity among groups.

Finally the minor value of variation among populations within groups was 0.21%, when tested Mediterranean Europe vs North of Europe and the US Caucasian vs East Asian vs Western Asia. This value permits to infer that formed groups were more homogeneous and also that the US Caucasian population is more similar to the countries from the North of Europe.

The small differentiation found between populations within groups support the present forensic settings that grouping reference populations into broad ethnic categories, when autosomal data are used [104]. This is, all the Caucasian population tested demonstrated to be very similar, even if located in different continents, like USA and Europe, and so a general database can be used for all Caucasian population.

Thus, the obtained results confirm that these 21 loci show an overall homogeneous distribution across European population, and so it is permitted to calculate European and US Caucasian cases with these allele frequencies, however, Asian countries like Korea, when possible, should not be calculated with this database, since it will overestimate the probability of paternity and power of exclusion.

5 Concluding Remarks

Nowadays, before the implementation of new techniques or methodologies an internal validation study must be performed. The results from these tests are crucial to understand the performance, limitations and potential of new methods, applied in a particular lab with their specific procedures and equipment. Through the internal validation process performed with the GlobalFiler™ Express kit, in SGBF-S with the South Portuguese population, all the loci have proven to be highly specific for human DNA and the results revealed to be robust and reproducible. The studies evidenced that the optimal parameters for PCR amplification for bloodstain samples were when using 22 cycles, with an input of DNA raging between 2 to 1 ng/µL and the threshold set on 100 RFUs. In the presence of problematic samples, like mixtures or degraded DNA, the threshold should be decreased to 50 RFUs and a more careful analysis must be performed to avoid loss of valuable information.

Full concordance between the 16 loci shared with Amp*Fl*STR® Identifiler Plus (Applied Biosystems) and PowerPlex® 16 HS (Promega Corporation) was demonstrated.

The sensitivity test demonstrated that GlobalFiler™ Express has less sensible when compared to other commercial kits, namely GlobalFiler™ that was designed for casework samples. However, reference samples have no problems of input quantity of DNA, and so, the low sensitivity will not be an obstacle for this kit's implementation.

Even though the method was designed for single source samples, throughout the contamination study, the presence of more than one contributor was always detected, by the observation of more than two peaks per locus, even in a ratio of 1:20, demonstrating that this method can distinguish single source samples from contaminated ones.

By studying the South Portuguese population, it was confirmed that all loci are in Hardy-Weinberg. As expected, the comparison study revealed that these markers show an overall homogeneous distribution across European

populations, being only statistically different from the Korean sample (Asian population). The AMOVA tests confirmed that populations living in the same continent tend to be similar to each other and show that the US Caucasian population is very similar to European populations.

It's important to note that, there are currently a large number of immigrants living in Portugal and it's common to have routine casework involving other nationalities. With those F_{ST} and AMOVA results, the study proved that these allele frequencies can be used to calculate correctly European and US Caucasian cases and, when needed, data exchange throughout Europe can be correctly accomplished. However, Asian population cases should not be calculated with this database, because the probability of paternity and power of exclusion will be overestimated. In the future, an analysis should be performed using GlobalFiler™ Express in the African population living in South Portugal. This is one of major groups of immigrants that appear in the service and, due the fact that any comparison was made in this study with African people, it's necessary to have a correct database for this population before the kit's implementation.

As predictable, the SE33 marker reveals to be the most polymorphic locus and TPOX the least one. The combined power of discrimination for the 21 autosomal loci was of 0.99999999999999999999999981765, the combined probability of match was 1.8356×10^{-26} and the combined power of exclusion was 0.99999999966339800. Those values, illustrate the improved performance of GlobalFiler™ Express when compared with Amp*Fl*STR ® Identifiler Plus (Applied Biosystems) and PowerPlex® 16 HS (Promega Corporation), that are the kits currently use in this laboratory routine.

Beside the improvement in forensic parameters, the large number of loci included in GlobalFiler™ Express will reduce the number of amplifications required to obtain the total loci needed for exchange data across different countries, this also saves time in the process, considering that a direct amplification takes less than 40 minutes and a complete profile is obtained in approximately 90 minutes.

In summary, our results prove that GlobalFiler™ Express is reliable to be used for forensic identification in South Portuguese population. This kit fulfills all

the requirements needed to use hereafter in routine casework and their implementation will be a valuable new tool in this forensic genetic lab.

6 References

[1] A.R.W. Jackson, J. J.M., Forensic Science, Third edit, Englan, 2011.

[2] W. Goodwin, A. Linacre, S. Hadi, An Introduction to Forensic Genetics, First edit, England, 2007.

[3] J.M. Butler, Forensic DNA typing: biology, technology, and genetics of STRs markers, Second edit, USA, 2005.

[4] M.F. Pinheiro, Algumas perspectivas da identificação genética, in: Genética Forense - Perspectivas da Identificação Genética, Pessoa, Ed, Porto, 2010: pp. 17 – 78.

[5] M. Jobling, P. Gill, Encoded evidence: DNA in forensic analysis., Nat. Rev. Genet. 5 (2004) 739–751.

[6] M. Sjerps, A. D. Kloosterman, Statistical aspects of interpreting DNA profiling in legal cases, Stat. Neerl. 57 (2003) 368–389.

[7] J.M. Butler, C.R. Hill, M.D. Coble, Variability of New STR Loci and Kits in US Population Groups Variability of New STR Loci and Kits in US Population Groups, Promega. (2012) 1–28.

[8] R. Fourney, K. Bowen, J. Elliott, Forensic reality and the practical experience of DNA typing update, DNA Data Bank (2002) 1–27.

[9] P. Thanakiatkrai, T. Kitpipit, Current STR-based techniques in forensic science, Maejo Int. J. Sci. Technol. 7 (2013) 1–15.

[10] P. Schneider, Basic issues in forensic DNA typing, Forensic Sci. Int. 88 (1997) 17–22.

[11] A.J. Jeffreys, Genetic fingerprinting, Nat. Med. 11 (2005) 1035–1039.

[12] R. Nussbaum, R.R. McInnes, H.F. Willard., Thompson & Tompson Genetics in Medicine, Senen, Elsevier Health Sciences, 2007.

[13] J. Pasternak, Human Population Genetics, Wiley-Blackwell, 2012.

[14] C.R. Hill, M.C. Kline, M.D. Coble, J.M. Butler, Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples., J. Forensic Sci. 53 (2008) 73–80.

[15] P. Gill, A. Jeffreys, D. Werrett, Forensic application of DNA'fingerprints', Nature. 318 (1985).

[16] A. Jelfreys, V. Wilson, S. Thein, Individual-specific "fingerprints" of human DNA, Nature. (1985).

[17] B. Alberts, D. Bray, K. Hopkins, A. Johnson, J. Lewis, M. Raff, et al., Essential Cell Biology, 2009.

[18] R. Decorte, Genetic identification in the 21st century - current status and future developments, Forensic Sci. Int. 201 (2010) 160–164.

[19] K. Mullis, F. Faloona, S. Scharf, Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction, Biotechnol. …. (1992).

[20] http://www.stamboomdeleeuw.nl/dnaengenealogie.html, (24.05.2014).

[21] Y. Zhong, B. Budowle, F. Center, The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems, Electrophoresis. 20 (1999) 1682–1696.

[22] J. Buckleton, C.M. Triggs, Relatedness and DNA: are we taking it seriously enough?, Forensic Sci. Int. 152 (2005) 115–9.

[23] B. Pierce, Genetics: A conceptual approach, Second Edit, 2005.

[24] T. Senge, B. Madea, A. Junge, M.A. Rothschild, P.M. Schneider, STRs, mini STRs and SNPs - a comparative study for typing degraded DNA., Leg. Med. Tokyo Japan. 13 (2011) 68–74.

[25] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology., Nat. Rev. Genet. 12 (2011) 179–92.

[26] P.A. da Costa Francez, E.M.R. Rodrigues, A.M. de Velasco, S.E.B. dos Santos, Insertion-deletion polymorphisms--utilization on forensic analysis., Int. J. Legal Med. 126 (2012) 491–6.

[27] R. Pereira, C. Phillips, C. Alves, A. Amorim, A. Carracedo, L. Gusmão, A new multiplex for human identification using insertion/deletion polymorphisms., Electrophoresis. 30 (2009) 3682–90.

[28] J.L. Weber, D. David, J. Heil, Y. Fan, C. Zhao, G. Marth, Human diallelic insertion/deletion polymorphisms., Am. J. Hum. Genet. 71 (2002) 854–862.

[29] http://blog.hackbrightacademy.com/2013/07/indel-finder-how-the-python-version-of-this-program-works/, (24.05.2014).

[30] J.M. Butler, Advanced Topics in Forensic DNA Typing, USA, 2012.

[31] R. Szibor, M. Krawczak, S. Hering, J. Edelmann, E. Kuhlisch, D. Krause, Use of X-linked markers for forensic purposes., Int. J. Legal Med. 117 (2003) 67–74.

[32] J. Butler, Recent developments in Y-short tandem repeat and Y-single nucleotide polymorphism analysis, Forensic Sci. Rev. 15 (2003) 91–111.

[33] J. Butler, Fundamentals of forensic DNA typing, USA, 2009.

[34] A. Edwards, A. Civitello, H.A. Hammond, T. Caskey, DNA typing and genetic mapping with trimeric and tetrameric tandem repeats., Am. J. Hum. Genet. 49 (1991) 746–56.

[35] S. Bell, Encyclopedia of forensic science, 2009.

[36] A.J. Jeffreys, M.J. Allen, E. Hagelberg, A. Sonnberg, Identification of the skeletal remains of Mengele, Josef by DNA analysis, Forensic Sci. Int. 56 (1992) 65–76.

[37] K. Tamaki, A.J. Jeffreys, Human tandem repeat sequences in forensic DNA typing., Leg. Med. (Tokyo). 7 (2005) 244–50.

[38] J. Butler, C. Hill, Biology and genetics of new autosomal STR loci useful for forensic DNA analysis, Forensic Sci. Rev. 24 (2012) 15–26.

[39] P. Schneider, Expansion of the European Standard Set of DNA database loci-the current situation, Profiles DNA. (2009) 6–7.

[40] L. a Welch, P. Gill, C. Phillips, R. Ansell, N. Morling, W. Parson, et al., European Network of Forensic Science Institutes (ENFSI): Evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers., Forensic Sci. Int. Genet. 6 (2012) 819–26.

[41] J.M. Butler, Y. Shen, B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA., J. Forensic Sci. 48 (2003) 1054–64.

[42] E.A.M. Graham, Mini-STRs, Forensic Sci. Med. Pathol. 1 (2005) 65–66.

[43] A. Barbaro, L. Fernandez-Formoso, C. Phillips, a Carracedo, M. V Lareu, Casework application of a stand-alone pentaplex assay of extended-ESS STRs., Leg. Med. (Tokyo). 15 (2013) 217–21.

[44] P. Wiegand, M. Kleiber, Less is more--length reduction of STR amplicons using redesigned primers., Int. J. Legal Med. 114 (2001) 285–7.

[45] J. Butler, Short tandem repeat typing technologies used in human identity testing, Biotechniques. 43 (2007) Sii–Sv.

[46] X. Lin, J. Wu, H. Li, Z. Wang, J.-M. Lin, Determination of mini-short tandem repeat (miniSTR) loci by using the combination of polymerase chain reaction (PCR) and microchip electrophoresis., Talanta. 114 (2013) 131–7.

[47] J.M. Butler, MiniSTRs: Past, Present, and Future, Appl. Biosyst. Forensic News. 8311 (2006).

[48] B. Budowle, T.R. Moretti, K.M. Keys, B.W. Koons, J.B. Smerick, Validation studies of the CTT STR multiplex system., J. Forensic Sci. 42 (1997) 701–707.

[49] P. Gill, Role of short tandem repeat DNA in forensic casework in the UK-past, present, and future perspectives, Biotechniques. 385 (2002).

[50] Lifetechnologies Corporation, GlobalFiler ™ Express PCR Amplification Kit: User Guide, Applied Biosystems (2012).

[51] http://www.lifetechnologies.com/pt/en/home/industrial/human-identification/globalfiler-str-kit/powered-by-6-dye.html, (09.12.2013).

[52] M.L. Pontes, M.F. Pinheiro, Population data of the AmpFlSTR® NGM™ STR loci in a North of Portugal sample., Forensic Sci. Int. Genet. 6 (2012) e127–8.

[53] T. Ribeiro, P. Dario, N. Vital, S. Sanches, R. Espinheira, H. Geada, et al., Population data of the AmpFlSTR® NGM™ loci in South Portuguese population., Forensic Sci. Int. Genet. 7 (2013) e37–9.

[54] J.M. Butler, New Autosomal and Y-STR Loci and Kits: Making Data Driven Decisions Introdutory Remarks, in: ISHI Work. New Loci Kits, Atlanta, GA), 2013.

[55] J. Butler, Validation: What Is It, Why Does It Matter, and How Should It Be Done?, Appl. Biosyst. Forensic News. (2007).

[56] A. Methods, Scientific Working Group on DNA Analysis Methods Validation Guidelines for DNA Analysis Methods, (2012) 1–13.

[57] B. Beiguelman, S. Editora, Genética de populações humanas, Ribeirão Preto SBG. (2008).

[58] I. Evett, J. Buckleton, Statistical analysis of STR data, 16th Congr. Int. Soc. (1996) 14–15.

[59] http://www.pordata.pt/Portugal, (09.12.2013).

[60] http://mapa-de-portugal.blogspot.pt/2013_05_01_archive.html, (05.04.2014).

[61] David Levinson, Portugal, in: Greenwood (Ed.), Ethn. Groups Worldw. A Ready Ref. HandbooNo Title, Greenwood, 1998: p. 436.

[62] H. V. Livermore, A New History of Portugal, CUP Archive, 1969.

[63] David Levinson, Portugal, in: Greenwood (Ed.), Ethn. Groups Worldw. A Ready Ref. HandbooNo Title, Greenwood, 1998: p. 436.

[64] James Maxwell Anderson, The History of Portugal, Greenwood Publishing Group, 2000.

[65] A. Amorim, L. Gusmão, C. Alves, STR data (AmpFlSTR profiler plus) from north Portugal., Forensic Sci. Int. 115 (2001) 119–21.

[66] A.T. Fernandes, A. Brehm, Population data of five STRs in three regions from Portugal., Forensic Sci. Int. 129 (2002) 72–4.

[67] A. Fernandes, A. Brehm, C. Alves, L. Gusmao, A. Amorim, Genetic profile of the Madeira Archipelago population using the new PowerPlex® 16 System kit, Forensic Sci. Int. 125 (2002) 281–283.

[68] M.F. Pinheiro, L. Cainé, L. Pontes, D. Abrantes, G. Lima, M.J. Pereira, et al., Allele frequencies of sixteen STRs in the population of Northern Portugal., Forensic Sci. Int. 148 (2005) 221–3.

[69] L. Excoffier, G. Laval, S. Schneider, Arlequin (version 3.0): an integrated software package for population genetics data analysis., Evol. Bioinform. Online. 1 (2005) 47–50.

[70] A. Tereba, Tools for analysis of population statistics, Profiles in DNA 3, Promega Corporation, 1999, (n.d.).

[71] http://www.nfstc.org/pdi/Subject07/pdi_s07_m02_06_a.htm, (01.07.2014).

[72] J.R. Gilder, T.E. Doom, K. Inman, D.E. Krane, Run-specific limits of detection and quantitation for STR-based DNA testing., J. Forensic Sci. 52 (2007) 97–101.

[73] P. Gill, L. Fereday, N. Morling, P.M. Schneider, The evolution of DNA databases--recommendations for new European STR loci., Forensic Sci. Int. 156 (2006) 242–4.

[74] T.R. Moretti, a L. Baumstark, D. a Defenbaugh, K.M. Keys, J.B. Smerick, B. Budowle, Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples., J. Forensic Sci. 46 (2001) 647–60.

[75] Lifetechnologies Corporation, GlobalFiler™ PCR Amplification Kit: User Guide, Applied Biosystems (2012).

[76] C.R. Hill, M.C. Kline, J.J. Mulero, R.E. Lagac, C.W. Chang, L.K. Hennessy, et al., Concordance study between the AmpFlSTR® MiniFiler™ PCR amplification kit and conventional STR typing kits, J. Forensic Sci. 52 (2007) 870–873.

[77] http://www.cstl.nist.gov/strbase/var_SE33.htm, (24.07.2014).

[78] http://www.cstl.nist.gov/strbase/var_D1S1656.htm, (24.07.2014).

[79] C. Alves, V. Gomes, M.J. Prata, A. Amorim, L. Gusmão, Population data for Y-chromosome haplotypes defined by 17 STRs (AmpFlSTR YFiler) in Portugal., Forensic Sci. Int. 171 (2007) 250–5.

[80] M.L. Pontes, L. Cainé, D. Abrantes, G. Lima, M.F. Pinheiro, Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR Y-filer) in a Northern Portuguese population sample., Forensic Sci. Int. 170 (2007) 62–7.

[81] A. Jonkisz, B. Bartnik, T. Dobosz, Y-chromosomal polymorphic loci DYS19, DYS389 I/II, DYS390, DYS391, DYS392 and DYS393 in a population sample from southwestern Poland, Int. Congr. Ser. 1239 (2003) 353–355.

[82] N. Di Nunno, S.L. Baldassarra, C. Di Nunno, B. Boninfante, G. Guanti, G. Forleo, et al., Distribution of DYS391, DYS392, DYS393, DYS385, alleles in a Southern Italian population sample., J. Forensic Sci. 47 (2002) 911.

[83] H. Geada, R.M. Brito, T. Ribeiro, R. Espinheira, Portuguese population and paternity investigation studies with a multiplex PCR--the AmpFlSTR Profiler Plus., Forensic Sci. Int. 108 (2000) 31–7.

[84] C. Cruz, C. Vieira-Silva, T. Ribeiro, R. Espinheira, Genetic data for the locus SE33 in a south Portuguese population with Powerplex® ES System, Int. Congr. Ser. 1288 (2006) 427–429.

[85] A. Urquhart, N. Oldroyd, T. Downes, Selection of STR loci for forensic identification systems, … Soc. Forensic …. (1996).

[86] V. Lopes, A. Serra, J. Gamero, L. Sampaio, F. Balsa, C. Oliveira, et al., Allelic frequency distribution of 17 STRs from Identifiler and PowerPlex-16 in Central Portugal area and the Azores archipelago., Forensic Sci. Int. Genet. 4 (2009) e1–7.

[87] A. M. Bento, A. Semo, V. Lopes, V. Bogas, P. Brito, a. Serra, et al., Population data for Central Portugal population with NGM amplification kit, Forensic Sci. Int. Genet. Suppl. Ser. 4 (2013) e152–e153.

[88] O. García, J. Alonso, J. a Cano, R. García, G.M. Luque, P. Martín, et al., Population genetic data and concordance study for the kits Identifiler, NGM, PowerPlex ESX 17 System and Investigator ESSplex in Spain., Forensic Sci. Int. Genet. 6 (2012) e78–9.

[89] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci., Forensic Sci. Int. Genet. 7 (2013) e82–3.

[90] S.B. Hong, S.H. Kim, K.C. Kim, M.H. Park, J.Y. Lee, J.M. Song, et al., Korean population genetic data and concordance for the PowerPlex® ESX 17, AmpFlSTR Identifiler®, and PowerPlex® 16 systems., Forensic Sci. Int. Genet. 7 (2013) e47–51.

[91] A. El Andari, H. Othman, F. Taroni, I. Mansour, Population genetic data for 23 STR markers from Lebanon., Forensic Sci. Int. Genet. 7 (2013) e108–13.

[92] A. a Westen, T. Kraaijenbrink, E. a Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, et al., Comparing six commercial autosomal STR kits in a large Dutch population sample., Forensic Sci. Int. Genet. 10 (2014) 55–63.

[93] P. Hatzer-Grubwieser, B. Berger, D. Niederwieser, M. Steinlechner, Allele frequencies and concordance study of 16 STR loci-including the new European Standard Set (ESS) loci--in an Austrian population sample., Forensic Sci. Int. Genet. 6 (2012) e50–1.

[94] J. Ross, W. Parson, I. Furać, M. Kubat, M. Holland, Multiplex PCR amplification of eight STR loci in Austrian and Croatian Caucasian populations., Int. J. Legal Med. 115 (2001) 57–60.

[95] K. Montelius, A.O. Karlsson, G. Holmlund, STR data for the AmpFlSTR Identifiler loci from Swedish population in comparison to European, as well as with non-European population., Forensic Sci. Int. Genet. 2 (2008) e49–52.

[96] A.O. Tillmar, H. Nilsson, D. Kling, K. Montelius, Analysis of Investigator HDplex markers in Swedish and Somali populations., Forensic Sci. Int. Genet. 7 (2013) e21–2.

[97] L. Albinsson, L. Norén, R. Hedell, R. Ansell, Swedish population data and concordance for the kits PowerPlex® ESX 16 System, PowerPlex® ESI 16 System, AmpFlSTR® NGM™, AmpFlSTR® SGM Plus™ and Investigator ESSplex., Forensic Sci. Int. Genet. 5 (2011) e89–92.

[98] I. Ciuna, M. Guarnaccia, E. Ginestra, a. Agodi, D. Piscitello, S. Spitaleri, et al., Allele frequencies for STR loci in a Sicilian population: Genetic prevalence and disequilibrium, Int. Congr. Ser. 1288 (2006) 343–345.

[99] A. González-Liñán, L. Trizzino, D. Giambelluca, S. Salvo, A. Marino, A. Allegra, Population data of the 16 autosomal STRs loci of the Powerplex ESI 17 System in Sicilian population (Italy)., Forensic Sci. Int. Genet. 7 (2013) e93–4.

[100] P.G. Meirmans, P.W. Hedrick, Assessing population structure: F(ST) and related measures., Mol. Ecol. Resour. 11 (2011) 5–18.

[101] M. Roesti, W. Salzburger, D. Berner, Uninformative polymorphisms bias genome scans for signatures of selection., BMC Evol. Biol. 12 (2012) 94.

[102] E. Chouery, M.D. Coble, K.M. Strouss, J.L. Saunier, N. Jalkh, M. Medlej-Hashim, et al., Population genetic data for 17 STR markers from Lebanon., Leg. Med. (Tokyo). 12 (2010) 324–6.

[103] A. Berti, F. Brisighelli, A. Bosetti, E. Pilli, C. Trapani, V. Tullio, et al., Allele frequencies of the new European Standard Set (ESS) loci in the Italian population., Forensic Sci. Int. Genet. 5 (2011) 548–9.

[104] L.B. Jorde, W.S. Watkins, M.J. Bamshad, M.E. Dixon, C.E. Ricker, M.T. Seielstad, et al., The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data., Am. J. Hum. Genet. 66 (2000) 979–88.